UNIVERSITY OF TRENTO
DEPARTMENT OF PSYCHOLOGY AND COGNITIVE SCIENCES

DOCTORAL SCHOOL IN
PSYCHOLOGICAL SCIENCES AND EDUCATION

Doctoral Dissertation
in the subject of Cognitive Sciences

Expectations of Obedience and the Development of Moral Reasoning

Advisor
Prof. Luca Surian

PhD candidate
Francesco Margoni

October 2016

Contents

GENERAL INTRODUCTION

A Functional and Relational Perspective on Morality

A Functional and Relational Perspective on Morality

In order to understand moral judgment and moral behavior from a psychological standpoint, we should carefully consider the *function* of morality, and that, as a phenomenon, morality is fundamentally *relational*. This statement is not trivial as may appear to be, and here I insist on some main implications of it. In what follows, I first give a few coordinates to delimit the 'essence' of morality. Second, I briefly review some recent evidence in developmental and evolutionary psychology suggesting that both prosocial and aggressive tendencies reflect important aspects of our socio-moral lives. Third, I argue that morality has evolved in humans' groups to serve the function of maintaining prosocial tendencies and constraining aggressive ones, and moral ideals represent the actual balance between different forces and individuals who are necessarily in a specific relationship. Within this view, the obedience to authority and the acceptance of the hierarchical structure of society, along with the internalization of basic principles of conduct, represent core aspects of morality. Finally, I discuss the distinction between 'autonomy' and 'heteronomy' in moral development (Piaget, 1932; Kohlberg, 1969). Two main stages, levels or forms of morality will then be dissociated, one that consists in the simple obedience to rules or authority's mandates, and the other one that consists in the internal motivation to comply either with the authority or with a set of principles according to which we aim to live a morally good life. I argue that some aspects of heteronomy (that is, the compliance with authority's orders) remain, also in adulthood, an important part of our morality.

The aim of this introduction is to discuss some core aspects of morality, also by identifying the selective pressures that likely affected their evolution. This is done in order to lay a basis for the understanding of the content of the present dissertation, that mainly addresses questions regarding the development of core aspects of morality during infancy and childhood, namely (a) the ability to represent social power asymmetries and authority, and (b)

the ability to weigh agents' intentions in producing a moral judgment. The latter ability is linked with the child's capacity to evaluate not only whether the actions outcomes represent a violation of some rule or authority's mandate, but also whether the individual possess a mindset that will likely produce desirable or undesirable outcomes in the future.

1. The Relational Approach to the Definition of Morality

When studying the developmental aspects of humans' morality, one may feel the need to possess at least a rough or working definition of the phenomenon under consideration. Here, I do not offer a clear-cut definition of morality. Instead, I insist on some core aspects that would hopefully clarify the concept of morality and, as a result, it would help us to understand which psychological aspects we should necessarily consider when we engage in the endeavor of studying morality.

As suggested by the relational models theory (Fiske, 1991, 1992; but see also the chapter on Ethics in Hegel, 1807/1977) and its recent extension (Rai & Fiske, 2011), ethics is necessarily defined by the particular relationship that is in place between a certain set of individuals who share a specific spatiotemporal context. For example, the morality within a family setting can be dissociated from the morality within a group of pairs playing cards together, and the latter is again different from the morality of a citizen of a democratic European society. An individual who is a father could be also a card player, and could live in a democratic society; however, when he is parenting his children he would likely endorse a particular role and specific values that he does not necessarily endorse or prioritize in other social situations. He would act more authoritatively with his children than with his cards mates.

Of course, we do not need to posit as many moral faculties as the possible roles people may have in the society. Though, it is important to acknowledge that the 'moral faculty' is necessarily inserted in a specific social context, and it likely works to serve specific functions within that context. Depending on how the relevant social relation is construed, actions that we would typically judge as morally bad (like, for example, harming or killing an innocent victim) could be perceived as morally worth, for example when such an action is accomplished because of an authority's mandate or because the majority of people wants so (Fiske & Rai, 2014; Rai & Fiske, 2011).

1.1. Conformity and Obedience to Authority as Pillars of Morality

Social relationships are often part of a broader social context, where power asymmetries exist and hierarchies are formed to handle and minimize conflicts among group members, and, possibly, to manage between-group competitions. Within this frame, two main elements are essential for every possible morality in a certain context. A first pillar is the acceptance and the conformity to a minimum set of rules; as Piaget wrote in his 1932 influential book "all morality consists in a system of rules, and the essence of all morality is to be sought for in the respect which the individual acquires for these rules." A second pillar is the obedience to the established authority and the acceptance of the hierarchies.

It might seem somewhat 'Tory' to posit that the obedience to authority's rules and the acceptance of power asymmetries are the basics of morality, and I hypothesize that this feeling is related to our Western and democratic culture—citizens being educated (both formally and informally) to think accordingly to an individualistic view. Especially after the Cold War, but soon after the rise of modern state, people tended to resize the power of the state by putting the individual (with its own set of rights) before the community. However, we do not have to forget that each society, political system or morality *de facto* implies that a

number of individuals with different forces and subsequent positions within the system try to live together as a group, seeking the best balance between everyone force and needs.

Traditionally, Western developmental psychologists have generally agreed upon the view that morality evolves from an initial stage or phase where the child identifies the good with the obedience to an authority that remains external to his or her conscience, to an adult-like phase where morality is now internalized, and the respect for the rules is strictly related to the mutual respect and the cooperation between individuals with equal rights. The individual understands what are the reasons why society introduced the rule, and she or he understands that rules and authorities (should) protect some abstract and generally valid principles of justice and right (Kohlberg, 1969; Piaget, 1932).

Moreover, 'domain theorist' insisted in dissociating moral norms from social ones (Turiel, 1978; Nucci, 1981; Killen & Smetana, 2015). According to their view, moral rules reflect abstract and universally valid principles, and therefore do not depend upon any authority's mandate. By contrast, the prescriptive force of conventional norms depends entirely on the social context or the authority. A host of studies reported that, by at least the age of three, children can distinguish between moral and conventional rules, and consequently they judge moral violations—harming an innocent victim—more harshly than conventional transgressions, like wearing pajamas at school (e.g., Nucci, 1985; Smetana & Braeges, 1990). Moreover, conventional but not moral rules are judged to be authority-dependent.

1.2. Reacting to Classical Views

A solid reaction to the 'classical' views we mentioned has mainly come from studies that analyzed the moral judgment and reasoning of non-Western populations. Overall, these studies showed that some non-Western populations (e.g., Indian Brahmans and

'Untouchable', who live in a society where the moral code of "community" prevails over the Western-like moral code of "autonomy"; see Shweder, Much, Mahapatra, & Park, 1997) judge some conventional transgressions as they were moral (Haidt, Koller, & Dias, 1993; Nichols, 2004; Nisan, 1987; Shweder, Mahapatra, & Miller, 1987; but see also Astuti & Bloch, 2015). For example, participants were asked to judge violations of cultural norms like "The day after his father's death, the eldest son had a haircut and ate chicken". Indians judged those actions wrong, serious, unalterable and universally binding. That is, violations that do not imply harm were judged as they were moral violations. Moreover, there is evidence that some moral transgressions are judged also by Western adults as they were conventional violations (Kelly, Stich, Haley, Eng, & Fessler, 2007). Both these results undermine the view that the moral-conventional distinction is universal.

Furthermore, the universal validity of the stage-like model of moral development has also been criticized. For example, a recent study reported that adults living in small-scale societies, ranging from hunter-gatherer to pastoralist to horticulturalist, compared to Western individuals, when presented with a moral judgment task, take less into account the agents' intentions and more the actions outcomes (Barrett et al., 2016). This finding is important because it raises the suspect that at least some core aspects of morality (as developmental psychologists have traditionally conceived morality) may be indeed the product of our culture instead of the product of our phylogeny. Therefore, the emphasis that now researchers give to individual rights—e.g., see the Smetana (2006)'s definition of moral transgressions as acts that are "wrong because they have intrinsic effects for others' right and welfare", p. 121—may actually sway the current understanding of the phenomenon. Community-like values, such as the respect for authority and social hierarchies, or the respect for the common goods, may have been central aspects of our morality and likely still play an essential but somewhat undetected role in shaping our current morality (e.g., Graham, Haidt, & Nosek, 2009).

2. The Functional Approach to the Definition of Morality

Now, not all rules are moral. To give an example, we can differentiate between grammatical rules and moral rules. Both have normative power, but only the transgressions of the latter elicit in us what we are used to call a moral condemnation. How then do we distinguish between moral and amoral rules? The approach I follow here could be named *functional*. Moral rules do not have a special nature or reality. However, they serve particular functions, and by specifying those functions, we should be able to define morality.

2.1. The Concept of 'Balance' in the Moral Domain

Overall, morality can be conceptualized as the attainment of the right balance between the forces in the field. If we consider that each individual in a society exerts some power over others (e.g., he claims the right to work, in a modern society, or he aims to achieve a higher reproductive status, in a ancient society and perhaps still today etc.), we also see that a specific balance between forces would be in place in each context, and morality is either the acceptance of this balance or the ideal balance towards which our actions and society structures should approximate. As Stoic philosophers already suggested, morality is the acceptance of the right or logical order between individuals in a relationship. Following this line of reasoning, then, there is no morality outside a specific context or relationship, and morality could be always defined by the best balance between individual forces.

Take for a moment the well-known distinction in philosophy and in moral psychology between utilitarianism and deontology. According to utilitarianism, the morality of an action depends on its outcomes; an action is good when it maximizes the common good. By contrast, according to deontology, the morality of an action depends on whether the action is

consistent with some abstract moral principles or norms. A question central to moral philosophy is which view has to be preferred, while a question central to moral psychology is whether people reason accordingly to utilitarian or deontological principles, or both. However, what is clear to me is that both views may act in our society as distinct forces. Different individuals will hold different moral principles, and the same individual may be willing to switch his or her view depending on the particular context. According to our definition of morality, we should consider that a compromise between the two positions will likely occurs, as a results of a conflict between different interests, preferences and attitudes. Ultimately, there should be a right balance between different forces, those who favor utilitarianism and those who favor deontology, and this balance, that is the product of both moral factors and factors external to the moral domain, is what we need to recognize as our morality.

2.2. Hierarchy and Morality

However, I have not yet specified which criteria we should use to address the question "best balance to pursue what aims?" This question is simply another way to ask which functions morality serves. Overall, morality is functional to the enhancement of prosocial tendencies and the constraining of egoistic and aggressive tendencies. From an evolutionary standpoint, moral tendencies may have been selected for ensuring long-term social-cooperative relationships between group members (Darwin, 1859/1982; Fiske, 1991; Joyce, 2006).

Human nature is neither intrinsically good nor bad. First, because it would be childlike to judge nature with the vocabulary of morality. Second, and more relevantly here, because humans show both a 'bright' and a 'dark' side, in the sense that we may claim that human nature typically goes both in the direction of some goals that we evaluate or interpret as

morally good and in the direction of some other goals that instead we consider morally bad. On the one hand, we may advocate that in the natural condition of mankind the individual tends to dominate aggressively and to prevail over others without any concern for their rights—the idea of individual right being actually absent in such an original state. By contrast, we may claim that human nature is fundamentally good, but it is actually corrupted by our 'evil' cultural, social or political systems. These two opposing claims are of course reported as clear and extreme views. However, evolutionary literature and, more recently, infants' cognition and developmental literature suggest that both tendencies are constitutive of the human nature and are displayed from a very young age. It has also been proposed that some core morally good and evil tendencies are indeed innate—being the product of our phylogeny (Bloom, 2013; Wynn, 2008).

Human beings have been evolved as a social and cooperative animal species (Cosmides & Tooby, 2013; Tomasello, 2014; Wilson, 2012). Already during early infancy, humans possess a whole set of abilities to represent the complex social world around them (Baillargeon, Scott, He, Sloane, Setoh, Jin, Wu, & Bian, 2015; Banaji & Gelman, 2013). In the first year of life, infants possess the ability to distinguish between prosocial and antisocial agents, they prefer helping agents over hindering or mean agents and also expect others to hold the same preferences (Choi & Luo, 2015; Hamlin, Wynn, & Bloom, 2007; Lee, Yun, Kim, & Song, 2015; Meristo & Surian, 2014). Moreover, infants show an early sense of fairness. They prefer agents that distribute the resources in a fairly way rather than in an unfairly way, and expect others to have the same preferences they hold (e.g., Geraci & Surian, 2011; Schmidt & Sommerville, 2011; Sloane, Baillargeon, & Premack, 2012). Further studies show that infants represent dominance hierarchies and conflicts over resources (Mascaro & Csibra, 2012; Pun, Birch, & Baron, 2016; Thomsen, Frankenhuis, Ingold-Smith, & Carey, 2011), and that they expect subordinates to obey to a leader or an

authority but not to a bully that simply hits the subordinates in order to gain their respect (Margoni, Baillargeon, & Surian, 2016).

A number of studies also reported that infants and toddlers spontaneously help others that are in need and often share valuable resources with them (e.g., Brownell, Iesue, Nichols, & Svetlova, 2013; Svetlova, Nichols, & Brownell, 2010; Warneken & Tomasello, 2006). Overall, these results suggest that, very early in life, humans are able to represent prosocial and antisocial actions; they prefer morally good and fair agents to mean and unfair agents; and they show altruistic tendencies. These mental abilities and action tendencies might have been functional in supporting the group cohesion and in maintaining the cooperation between group members across the evolutionary trajectory of our species.

However, along with a 'bright side', humans also possess a resolute will to conflict and compete for resources and often tend to dominate each other (Aureli & de Waal, 2000; Darwin, 1859/1982; Lasker, 1907). The desires to overcome others and to possess more resources than others characterize our psychology (Bloom, 2013). Samuel Johnson famously stated, "[…] no two people can be half an hour together, but one shall acquire an evident superiority over the other". Both these egoistic and aggressive tendencies can be controlled and constrained, and they have been, by establishing hierarchical social structures (Fiske, 1991; Heinrich & Gil-White, 2001, Sidanius & Pratto, 1999). At least after the introduction of the agriculture and the increase of the population size, a hierarchical structuring became the rule in humans' societies (Boehm, 1999; Knauft, 1991; Nye, 2008).

Hierarchical social structures with leader-followers relationships, then, may have been evolved to constrain egoistic and aggressive tendencies. At the same time, a rigid social structure and a leader facilitate and maintain cooperation within the group. A current hypothesis is that leaders may solve the problem of free-riders (King, Johnson, & Van Vugt, 2009). It is well-known that the presence of free-riders (individuals who benefit from others'

effort without contributing) can undermine cooperation. A proposed solution to free-riders was punishment. However, it is not clear how the punishment mechanism may have been evolved, considering that punishment is a costly act (e.g., can represent a danger for the punisher). A clever evolutionary solution may then be leadership. Leaders are often the individuals who bear the cost of punishment. They may be willing to do that in return for those privileges accorded to them. Further evidence from experimental psychology supports this line of reasoning by showing that high level of cooperation is indeed achieved in the situation in which only one individual is in charge of punishment (O'Gorman, Henrich, & Van Vugt, 2009).

To sum up, then, morality may have been evolved to support and maintain cohesion and collaboration between members of the same group, and to constrain aggressive and egoistic tendencies into coordinated actions that will pursue shared goals or maximize the group benefits in term of resources and capacity to defend the group against rival groups. Morality dictates the right balance between individual forces engaging in specific relationships, keeps in-group members together by enhancing fairness and collaboration, and helps in maintaining the order within the group—that again is functional to the cooperation maintenance—by constraining individuals to obey to current rules and to respect the authority's power and the social hierarchy.

3. Two Levels of Morality

The Kant (1788/2002)'s distinction between 'heteronomy of the power of choice' and 'autonomy of the will' was taken up again by Piaget (1932)'s work on the development of children's moral reasoning. According to Piaget—and, later on, Kohlberg—not only these

concepts are distinct and identify two opposing ways to approach morality, but they also characterize different stages in the development of moral reasoning.

In a first stage of the consciousness of the rules, the child submits completely in intention to the rules given by an adult authority, but these moral rules remain external to her or his conscience. This explains why the child considers the rules as sacred and, at the same time, rules do not really guide or transform her or his moral practice. The child is still in a stage characterized by a 'heteronomous' respect for the rules and a 'moral realism' or an 'objective' conception of responsibility, in which the child observes the letter more than the spirit of the law. With regard to the moral judgment, when asked to choose which character is the naughtiest and deserves to be punished—between a supposedly well-intentioned character that accidentally caused serious material damage and a bad-intentioned one that accidentally caused a less serious damage—children attend to outcomes and subsequently condemn the well-intentioned character more than the bad-intentioned one. This is because they focus on the moral violation and the letter of the law, regardless of the spirit of the moral rules and the true intentions behind the agent' action.

In a second stage of the consciousness of the rules, children develop the 'autonomy' of the conscience. Moral rules and adults' instructions are interiorized and generalized. The child now understands why the rules exist in the first place, that is, because of maintaining cooperation and order within the group. Children no longer consider rules as sacred. Instead, they consider rules as based on a general agreement between individuals. Now is the spirit of the law that matters, not the letter. Children's conception of responsibility shifts from 'objective' to 'subjective', and they judge the morality of an action or an agent attending to intentions rather than external outcomes.

According to recent evidence, the 'realist' outcome-based reasoning is typical in younger preschool children (e.g., Cushman et al., 2013; Margoni & Surian, 2017), but, as we

already discussed, to some extent both the autonomous and the heteronomous 'levels' are part of the adults' moral experience. Kant (1788/2002) argued that heteronomy of the power of choice "not only is no basis for any obligation at all but is, rather, opposed to the principle of obligation and to the morality of the will" (p. 48). However, the obedience to authority and, for example, the acceptance of the current laws of a State are *de facto* essential parts of everyone's morality—as it was clear to Descartes (1637/1970) when he presented his rudimentary morality.

It is often argued that conformity and obedience to authority are not *per se* part of the morality, also because these aspects of the human nature hide some dangerous pitfalls. At this point, everyone immediately visualizes the appalling events occurred to Jews during the Nazi regime. When the evilness of the actions is so clear, to condemn the Nazi hierarchs and the soldiers who obeyed to the commands become a very easy task. However, no society or large group of humans could survive or live in (relative) peace without conformity, respect for the rules (either internal or external), and obedience to authority. That is why we sometimes confront ourselves with the dilemma whether to obey an authority's mandate that we deem unfair or wrong. For example, should the soldier obey when his superiors order him to kill innocent people because this is the only means to hit a strategic target?

Milgram (1974) conducted a number of experiments and overall reported that people were willing to obey authority or experimenter's instructions also when they were asked to administer electric shocks that made helpless victims clearly suffering. Milgram and his collaborators remained puzzled by the findings. They ended up interpreting the participants' actions as determined by a conflict between external and constraining factors ("you should obey to authority") and participants' internal values ("you should not harm innocent others"). However, following our arguments, here we should rather claim that what occurred in Milgram's participants was a conflict *between values*—a conflict between the value of

preserving others' rights and the value, perhaps less available to consciousness but still working in the participants' mind, of obedience to authority. Then it should be very simple to address the question why morality, that wins over everything, does not win over authority—that is, because complying with the authority and respecting the rules are at the very core of our morality, whether we like it or we do not.

4. The Content of the Present Dissertation

The present dissertation collects several works that are either published or submitted to relevant journals in the field of developmental or experimental psychology. In the first chapter, I report three experiments overall suggesting that 21-month-old infants are already able to distinguish between *social dominance* (an asymmetry in which a dominant individual prevails over subordinates in competitive situations, typically by exert force or coercion) and *authority* or *leadership* (a social asymmetry in which the power of an authority over subordinates is deemed rightful by the parties involved). Infants saw geometric-shaped computer animations showing either a bully hitting a group of subordinates or a leader that respected the subordinates. Both dominant figures gave orders to the subordinates, and subordinates complied or violated their instructions. By using the *violation-of-expectation* (VOE) paradigm, we found that infants expect subordinates to comply with a leader's instruction but not with a bully's instruction, unless the bully closely controlled the subordinates. These findings suggest that even 21-month-olds possess an understanding of these complex dynamics of power, authority and obedience. I argue that this is an important building block in the acquisition of moral knowledge.

In the second chapter, I present two experiments conducted on children between ages four and eight. For the first time, we reported that during preschool years a shift occurs in

children's moral judgment of helping actions. Between ages 4 and 6, a crucial developmental change from an outcome-based to an intent-based goodness judgment occurs. The verbal judgments of children undergoes a shift from an 'objective' to a 'subjective' conception of responsibility. To rely on the agents' intentions rather than on actions outcomes is part of a mature way to produce a moral judgment. In chapter six, indeed, I report evidence suggesting that the variable of the agent's intention explains most of the variability of adults' moral judgments of cases of both help and harm.

A related research question, also partially addressed in chapter two, is whether the crucial outcome-to-intent shift occurring during preschool years reflects a conceptual change within the moral domain or ancillary changes occurring outside the moral domain, for example in executive functioning skills or in theory of mind. The data I present here tentatively support a 'continuous account', that is, the hypothesis that changes in moral judgment reflects changes occurring outside the moral domain, and that the (intent-based) moral concepts remain unchanged during the lifespan. I took up again this issue in chapter four, where I argue that changes in executive function are likely the best candidate to explain the current finding regarding the outcome-to-intent shift in moral judgment. Indeed, studies on infants' socio-moral evaluations, which used spontaneous-response tasks rather than elicited-response tasks, thus reducing the processing demands, show that, very early in life, humans evaluate others' actions taking into account the agents' intentions (e.g., Dunfield & Kuhlmeier, 2010; Hamlin, 2013; Lee et al., 2015). Therefore, it can be argued that developmental differences result from changes in executive functioning skills (or theory of mind skills, which however rely on executive function), and that the concept of moral goodness or badness remains unchanged throughout our life.

A continuity hypothesis is further supported by the finding presented in the seventh and last chapter. There I report a study on the difference between younger (age range: 21—39)

and older adults (age range: 63—90) in the use of intention and outcome during the production of a moral judgment. Participants rated the moral goodness or badness of helping and harming actions. Results show that older adults' moral evaluations rely less on agents' intentions and more on actions outcomes compared to the moral evaluations on the younger adults. Further analyses confirmed that this 'intention-to-outcome shift' taking place late in life can be explained by ancillary changes occurring outside the moral domain.

Then, in chapter three, two types of intention are dissociated, a biocentric intention (i.e., the agent preserves nature because of nature intrinsic value) and an anthropocentric intention (i.e., the agent preserves nature because it helps human's interests). The study aimed to investigate whether preschool children, who start judging the morality of actions based on an intention assessment, use this distinction between intention types in their moral evaluation and, as a result, judge agents that acted with a biocentric intention more bad (when causing an harm) or good (when helping others) than agents with anthropocentric intentions. The results show only an emerging preference for biocentrism. Therefore, intentions that are associated with different moral views (biocentrism and anthropocentrism) only partially affect the emerging intent-based judgment of preschool children.

Finally, in chapter five, I present a brief review of studies on mental state reasoning in the moral judgment of children and adults with autism spectrum disorder (ASD). This clinical population is known for having theory of mind impairments, thus it proves useful in understanding the role of factors external to the moral domain in determining the typical developmental trajectory of moral judgment. The aim of the review was to clarify whether ASD children develop the ability to judge the morality of an action or an agent by relying on the agent's intentions. Current evidence suggests that the impairment in theory of mind hinders the development of a moral judgment based on the agents' mental states in children with ASD. This clinical population shows a preserved capacity to produce a basic moral

judgment when evaluating those cases where, for example, an intention to violate a moral rule is followed by a negative consequence for the victim. However, this preserved capacity is explained by the fact that ASD children mostly evaluate the outcome or other irrelevant external factors such as the victim's emotional reaction. Indeed, when presented with ambiguous cases that require the analysis of the agent's mental states (e.g., the case of a failed attempt to harm), ASD individuals encounter some difficulties in producing an intent-based moral judgment.

In conclusion, in the present dissertation, I report new findings on some developmental aspects regarding both a first and core level of morality, i.e., the respect for rules and authority figures, and a second level of morality, i.e., the understanding of the importance of the spirit of the law and the use of the agents' mental states during the production of a moral judgment. I conclude that infants are already able to understand that individuals obey to instructions given by authorities, but not by bullies. Moreover, I conclude that the crucial period for children to show an intent-based judgment when presented with verbal tasks is about ages 5-6. However, I argue that this development in moral judgment reflects changes occurring outside the moral domain, so that changes in executive functioning skills determine the emergence of a capacity to evaluate based on an intention assessment that is likely to be already present during the early infancy. One future direction would be to collect evidence in order to investigate whether both levels of morality are based on domain-specific, innate mechanisms and indeed develop early in human life.

PART 1

INFANCY: EXPECTATIONS OF OBEDIENCE

CHAPTER 1

Infants Distinguish Between Dominance and Leadership

This chapter is based on the following original article:

Margoni, F., Baillargeon, R., & Surian, L. (2016). *Infants distinguish between leaders and bullies.* Manuscript submitted.

Abstract

Across three experiments, we investigated whether 21-month-olds distinguish between social dominance and leadership. In the former, the dominant prevails over the subordinates in competitive situations, whereas in the latter the power of the leader is deemed rightful by the parties involved. Infants were presented with computer animations and were familiarized with a bully that hit the subordinates or with a leader that did not hit subordinates. Both the bully and the leader then gave orders to the subordinates. During the test phase, infants saw the subordinates complying or violating the orders. In the absence of the dominant agent, infants expected subordinates to obey to the leader, but not to the bully. However, they expected the subordinates to obey when the bully did not leave the scene. These results suggest that the ability to represent different forms of social power asymmetries develops early in life.

Infants Distinguish Between Dominance and Leadership

Human beings are among the most social and cooperative animal species (Tomasello, 2014; Wilson, 2012). Our particular evolutionary trajectory has provided us with early-emerging mechanisms for representing others intentions, desires, beliefs and behaviors, along with an early-developing ability to infer complex social relations among in-group members (Fiske, 1991; Richerson & Boyd, 2006). Evidence of these early-emerging mental representational mechanisms come from a number of recent studies on infants' cognition (e.g., Baillargeon, Scott, He, Sloane, Setoh, Jin, Wu, & Bian, 2015; Bloom, 2013; Hamlin, Wynn, & Bloom, 2007; Surian, Caldi, & Sperber, 2007; Wynn, 2008).

Already in the first year of life, infants' social expectations rely on others intentions and they understand both successful and failed actions (Brandone & Wellman, 2009; Dunfield & Kuhlmeier, 2010; Gergely & Csibra, 2003; Hamlin, 2013; Lee, Yun, Kim, & Song, 2015; Margoni & Surian, 2016a; Woodward, 1998). In the second year of life, infants understand that others may hold false beliefs, suggesting that they already possess a theory of mind (Baillargeon, Scott, & He, 2010; Onishi & Baillargeon, 2005; Surian et al., 2007; Wellman, 2014).

Infants' theory of mind skills inform also their socio-moral expectations and evaluations. Within the first year of life, infants develop the ability to distinguish between prosocial and antisocial agents, show to prefer helping agents to hindering agents, and expect that others would hold these same moral preferences (Hamlin & Wynn, 2011; Hamlin et al., 2007; Hamlin, Wynn, & Bloom, 2010; Lee et al., 2015; Meristo & Surian, 2014). Infants also possess an early-emerging understanding of fairness, and prefer agents that distribute the resources fairly rather than unfairly (e.g., Geraci & Surian, 2011; Schmidt & Sommerville, 2011; Sloane, Baillargeon, & Premack, 2012).

Furthermore, studies on infants and toddlers' behaviors reported that very early in life humans show a spontaneous tendency to altruistically help others (Hepach, Haberl, Lambert, & Tomasello, 2016; Over & Carpenter, 2009; Svetlova, Nichols, & Brownell, 2010; Warneken & Tomasello, 2006) and to share resources with them (Brownell, Iesue, Nichols, & Svetlova, 2013). An evolutionary account of these early-emerging mental abilities and socio-moral behavioral tendencies would insist on their functional role in supporting group cohesion and maintaining the collaboration among in-group members (e.g., Baumard, André, & Sperber, 2013; Bloom, 2013; Greene, 2013; Tomasello, Melis, Tennie, Wyman, & Herrmann, 2012). Within this standpoint, social bonds and morality itself can be explained as functional to the pursuing of in-group interests, so that, as a result, today we still possess moralities and religions that tended to systematically exclude out-group members (Bloom, 2012; Greene, 2013; see also Andrighetto, Baldissarri, Lattanzio, Loughnan, & Volpato, 2014).

Together with the prosocial tendencies related to the function of maintaining social cohesion, human nature consists also in a fierce will to compete for resources within the group and, as a result, humans show to be characterized by a tendency to dominate each other (Aureli & de Waal, 2000; Darwin, 1859/1982; Hobbes, 1651/1982; Lasker, 1907; von Clausewitz, 1832/1984). Egoistic motives and aggressive behaviors are widely spread in social and non-social species, especially when individuals or groups compete for scarce resources.

However, egoistic and aggressive tendencies had been usually controlled and constrained by the development of hierarchical social structures (Berger, Rosenholtz, & Zelditch, 1980; Fiske, 1991; Heinrich & Gil-White, 2001; Maslow, 1936; Sidanius & Pratto, 1999; Silk, 2007). In humans, this may have happened after the increase of population size and the following possibility for large-group society to compete with each other (Tomasello, 2014).

A hierarchical social structure helps coordination and cooperation among in-group members in fundamental activities such as foraging, resource distribution and warfare (Axelrod & Hamilton, 1981; Fiske, 2010; Overbeck, 2010; Tooby & Cosmides, 1992). Therefore, subordinates' obedience to dominants, leaders or legitimate authorities plays a crucial role in generating and maintaining social relationships within the group, and determined the exit from the Hobbesian 'pure state of nature' (Hobbes, 1651/1982).

Infants' Representation of Social Dominance

A number of studies have been conducted on infants' representations of dominance relations. Ten-month-olds already rely on the agents' relative size to predict the outcome of a conflicting situation (Thomsen, Frankenhuis, Ingold-Smith, & Carey, 2011). When presented with events in which two agents, a bigger and a smaller one, block each other's path of motion, infants expect that the smaller agent will clear the path for the bigger agent. A further study revealed that by 6 months infants understand social dominance relationships and, in particular, they are able to rely on numerical group size to predict the outcome of a dominance relation (Pun, Birch, & Baron, 2016; see also Lourenco, Bonny, & Schwartz, 2016; Pietraszewski & Shaw, 2015). Six-month-olds expect that an agent from a numerically larger group will prevail over an agent from a numerical smaller group.

Further studies reported that infants possess a sophisticated understanding of social dominance relations, where dominance has been defined "the tendency to prevail when one's goals conflict with those of another agent" (Mascaro & Csibra, 2012; see also Dahl, 1957; Hand, 1986; Russell, 1938; Weber, 1946). Fifteen-month-olds expect that an asymmetric relation between a dominant agent and a subordinate will generalize to different time and situations. However, they do not expect an agent that have been shown to be dominant over a certain subordinate to be also dominant over a new agent, whose relation with the dominant is unclear. Therefore, already in the second year of life, infants represent social dominance as a

relation rather than a stable individual property (Mascaro & Csibra, 2012). Moreover, 15-month-olds have been shown to be able to represent also social dominance hierarchies with more than two individuals by combining incrementally representations of several dyadic relations (Mascaro & Csibra, 2014), and they can infer dominance hierarchies by using a transitive inference (Gazes, Hampton, & Lourenco, 2015). Overall, these results show that humans develop very early in life the ability to represent important hierarchical aspects of our social world.

Social Dominance vs. Leadership

In the evolutionary and adult literature, we find an important distinction between two forms of dominance (Cheng, Tracy, Foulsham, Kingstone, & Henrich, 2013; Fiske, 1992; Henrich & Gil-White, 2001; Hogan, Curphy, & Hogan, 1994; Milgram, 1974; Van Vugt, 2006; von Rueden, Gurven, & Kaplan, 2011). A simpler one, often referred as *dominance*, is the capacity to prevail in competitive contexts. A more complex one, referred as *leadership* or *prestige*, is a social asymmetry in which the leader's position and powers are deemed rightful or legitimate by the parties involved. Dominance relations are then handled by the direct use of force or by intimidation. As we have seen, infants in the first year of life already understand this simpler form of social asymmetry (Mascaro & Csibra, 2012; Pun et al., 2016; Thomsen et al., 2011).

However, evidence showed that in some cases adults use both dominant and prestige or leadership strategies to gain influence over others and attain social rank (Cheng et al., 2013), though they prefer to compete for status not by bullying but by increasing the group perception that they are competent, generous and committed to shared values and aims (Anderson & Kilduff, 2009). Indeed, groups often punish aggressive members that use force in order to gain influence over others, and conferred a higher status to competent individuals (Ridgeway & Diekeman, 1989).

Leadership or prestige relationships may have even evolved to serve a different function than aggressive or dominant behaviors. Prestige or leadership may have evolved from selection pressures to facilitate the imitation and cultural learning processes that permit group members to acquire knowledge from the most skilled or competent individuals (Boyd & Richerson, 1985; Laland & Galef, 2009). Leaders need to show a higher intelligence, competence, group commitment and prosociality, and they are followed because they are respected since they represent a value for the entire group (see Gebert, Heinitz, & Buengeler, 2016) and because subordinates can learn from them (Aidar, 1989; Berger, Cohen, & Zelditch, 1972). By contrast, pure dominant individuals are simply able to win physical conflicts, and when subordinates comply with their instructions is because of fear. Moreover, a study of a small-scale Amerindian society revealed that individuals that have a community-wide influence gain also more and different fitness payoffs compared to individuals that simply show to be able to win physical conflicts (von Rueden et al., 2011).

Further studies argued that the leader-follower relation is fundamental to human societies, but is different from the dominant-subordinate relation (King, Johnson, & Van Vugt, 2009; Van Vugt, 2006; von Rueden & Van Vugt, 2015; see also Gülgöz & Gelman, in press). No past or present human society is without leader-followers relations (Bass, 1990; Boehm, 1999; Brown, 1991; Diamond, 1998; Lewis, 1974). This is because leaders serve the fundamental function of generating and maintaining coordination between individuals to achieve shared goals such as the group defense or resources distribution (Guinote & Vescio, 2010; Van Vugt, 2006), and they help the culture transmission by allowing others to imitate their behaviors (Berger et al., 1972; Henrich & Gil-White, 2001). Consistently, Fiske (1992) posits that all cultures rely on an ‘authority ranking’ model, that is, people attend to their and others relative positions in some existing hierarchical social dimension and are motivated by hierarchy rather than fear of coercive power to comply with leaders’ instructions. Moreover,

leaders have both privileges and a duty to protect and care for subordinates (Fiske & Haslam, 2005; Rai & Fiske, 2011).

The Present Research

By relying on the distinction between social dominance and leadership, for the first time our study assessed whether infants expect subordinates to comply with a leader's instruction but not a bully's instruction. Previous work showed that infants are able to represent social dominance relations (Mascaro & Csibra, 2012; Pun et al., 2016; Thomsen et al., 2011), but to the best of our knowledge no studies assessed whether they are also able to represent leader-followers relations. We designed a violation-of-expectation task, presenting 21-month-olds with a bully or a leader agent, and we asked whether infants expect subordinates to obey to the dominant's order even after she left.

Specifically, in Experiment 1 infants were familiarized either with a bullying agent (that hit with a stick the subordinates and stole their ball) or, in a different condition, with a leader (here the subordinates bowed to her and spontaneously offered her the ball). In the test events of both conditions, the dominant agent ordered the subordinates, and they either obeyed or disobeyed after she left. If infants understand the difference between a bully and a leader, they should expect subordinates to comply with the leader's instructions but should hold no expectation about the subordinates' behaviors in the bully condition. Indeed, we reasoned that people feel no obligation to act as a bully says, and if they do, it is because of fear, but people spontaneously feel the obligation to act as a leader or a legitimate authority commands them to do. In Experiment 2, infants saw identical events to those showed in the bully condition except that now the bully did not leave when subordinates obeyed or disobeyed. We predicted that infants would now expect subordinates to obey since they fear the presence of the bully.

Lastly, in Experiment 3 infants were familiarized with an agent that acted friendly towards the three subordinates without, however, being characterized as a leader. With this experiment, we wanted to exclude the alternative hypothesis that infants in Leader condition of Experiment 1 expected subordinates to comply because of a general positive interaction between the figures, and not because of the dominant being characterized as a leader. We predicted that infants would hold no expectations or that they would expect subordinates to disobey because no power asymmetry was shown.

Finding that infants expect subordinates to comply with leader's instructions in her absence and with bully's instructions only when the dominant is present may constrain the current understanding of infants' socio-moral expectations, showing that an early-developing 'naïve sociology' (see Hirschfeld, 1999) includes the ability to understand the different ways in which individuals may exercise power and comply with it.

Experiment 1

Design. Infants were assigned to a leader or a bully condition. In the *leader* condition of Experiment 1, we investigated whether 21-month-olds expect that subordinates would comply with an order given by an authority. Infants watched computer animations of interacting geometrical figures, a yellow leader with a stick and a hat, and three red subordinates. In the character-familiarization event, infants watched three subordinates playing with a ball until the leader arrived; the subordinates bowed to her and she bowed in response saying "ohhh", the subordinates offered her the ball, and the leader left the scene. In the following instruction-familiarization event, infants watched the leader ordering the subordinates to go to bed by saying "time for bed" and pointing to their house with her stick. The infants received two character-familiarization trials and two more instruction-familiarization trials. In each character-familiarization trial, infants watched a repetition of a

maximum of four familiarization events. In each instruction-familiarization trial, infants watched a repetition of a maximum of six events.

In the test events, infants were presented with the leader that ordered subordinates to go to bed. In the disobedience event (D), the subordinates complied with the leader's order and entered in their small house while the leader watched, but disobeyed after she left. In the obedience event (O), the subordinates continued to comply also after the leader left; they entered in their house and they close their eyes. The infants received four test trials; half group received a D-O-D-O test order, while the other half received an O-D-O-D order. In each test trial, infants watched a repetition of a maximum of four events.

In the *bully* condition, infants were presented with identical events except that the leader was replaced by a yellow oval agent carrying a stick. In the character-familiarization event, the bully hit the subordinates with the stick and they said "auch, auch", then she stole their ball and left the scene. By adding the bully condition, we investigated whether 21-month-olds expect that subordinates would comply with a bully's order in her absence.

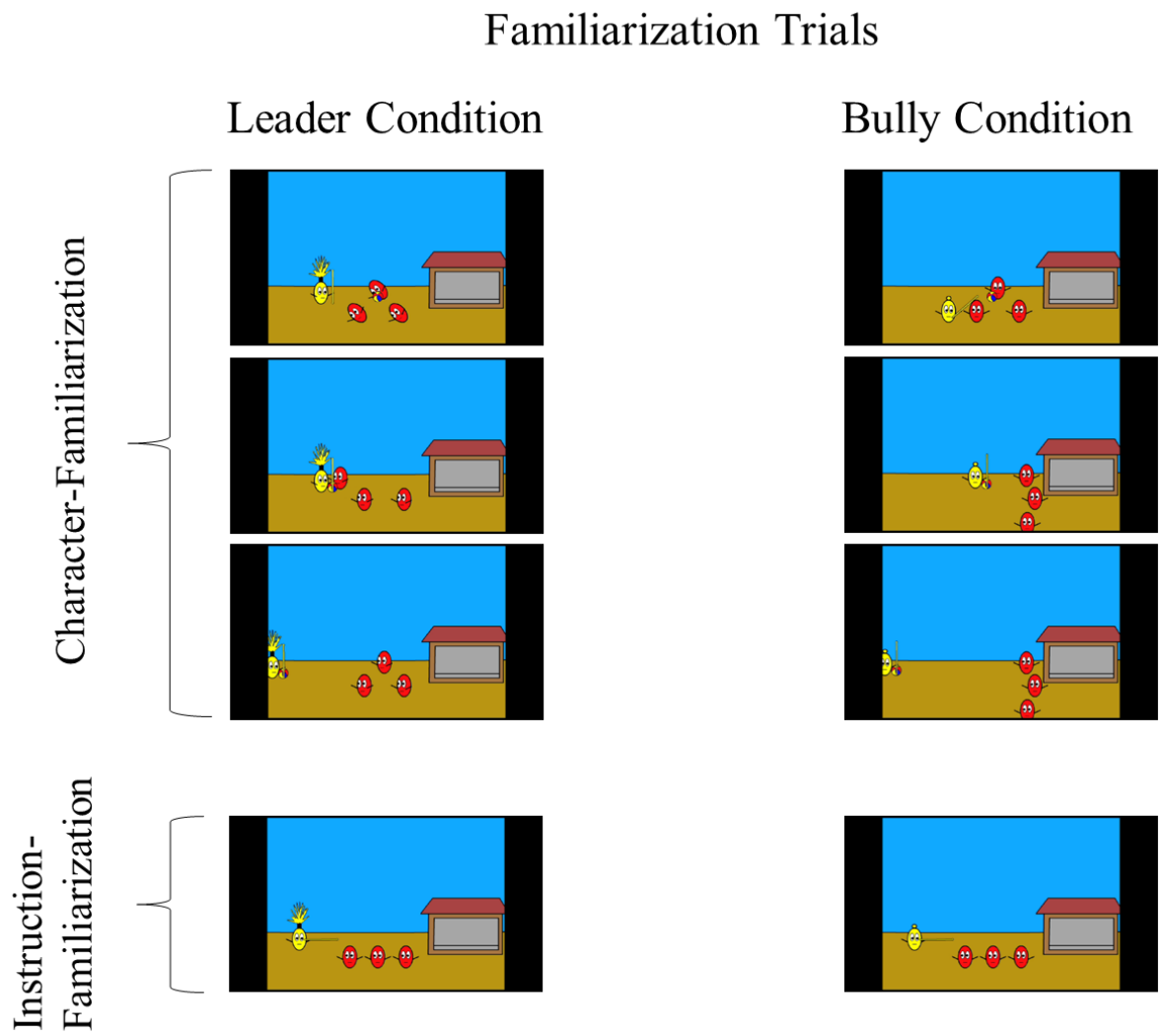

Figure 1. Schematic depiction of the events showed in the familiarization trials (character- and instruction-familiarizations) in leader (left side) and bully (right side) condition in Experiment 1. Each infant saw two character-familiarization trials and two instruction-familiarization trials.

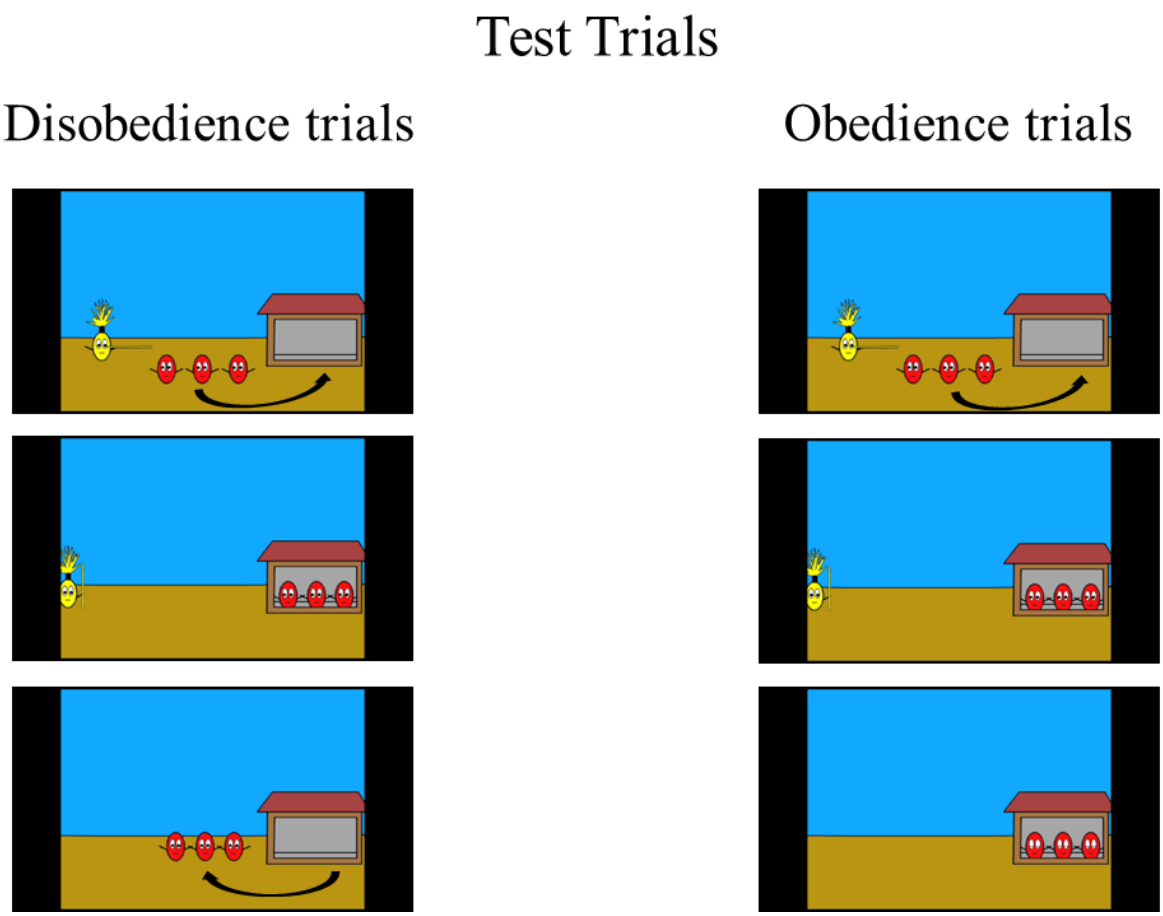

Figure 2. Schematic depiction of the events showed in the test trials (disobedience and obedience) in leader condition in Experiment 1. In bully condition, infants saw identical events except that a bully ordered to go to bed. Each infant saw four test trials.

Method.

Participants. Participants were 32 healthy full-term infants, 16 male (20 months, 1 day to 23 months, 22 days, *M* = 21 months, 15 days). An additional 12 infants were excluded due to fussiness (3), because they looked the maximum time allowed in all the test trials (7), or because they had a test looking time over 3 standard deviations from the mean (2). Equal numbers of infants were assigned to each condition (leader, bully). Infants' parents provided written informed consent, and the protocol was approved by the local Ethics Committee.

Materials and procedure. Infants sat individually on a parent's lap centered in front of an apparatus consisted of a display booth (201 cm high × 102 cm wide × 57 cm deep) with a large opening (46 cm × 95 cm) in its front wall where a projection screen (x cm × x cm) were inserted. Videos were projected onto the screen at eye-level to infants. Between trials, a supervisor lowered a curtain in front of this opening. Two cameras captured both the image of the events and the infants' looking behavior. Parents were instructed to remain silent and close their eyes during the test trial.

Each infant's looking behavior was monitored on-line by two naïve observers hidden on either side of the apparatus. The primary observer's responses were used in the analysis. Interobserver agreement was measured for 31/32 infants (only one observer was present for the other infant) and averaged 96% per trial per infants. Observers were blind to test events' order, and they guessed the right order at chance level (.47). In the Leader condition, interobserver agreement was measured for 16 infants, and averaged 96%; observers guessed the right order at chance level (.41). In the Bully condition, interobserver agreement was measured for 15/16 infants and averaged 96%; observers guessed the right order at chance level (.58).

Each trial began with an attention-getting still and smiling baby face. Each trial ended when the infant (a) looked away for 2 consecutive seconds after having looked for at least 25 (first two familiarization trials) or 7 (third and fourth familiarization trials) or 15 (test) cumulative seconds or (b) looked for a maximum of 75 (first two familiarization) or 35 (last two familiarization) or 60 (test) cumulative seconds.

Preliminary analyses revealed no significant interactions of condition and trial with infants' sex or number of siblings, all Fs < 1; the data were therefore collapsed across these factors in subsequent analyses.

Results.

Infants' looking times during test trials were analyzed performing a 2 × 2 mixed design ANOVA with trial (disobedience, obedience) as within-subjects factor and condition (bully, leader) as between-subjects factor (Fig. 3). We found only a significant Trial × Condition interaction, $F(1, 30) = 8.52$, $p = .007$. In the condition in which the leader was absent during the test phase (leader condition), infants looked reliably longer at the disobedience event ($M = 44.6$, $SD = 11.4$) than at the obedience event ($M = 34.9$, $SD = 11.8$), $t(15) = 3.77$, $p = .002$, $d = 1.01$ (two-tailed); in the bully condition, instead, infants looked equally longer at the two events, $p = .34$. The effect we found in the leader condition was due to a significant difference in the second pair of test events, $t(12) = 2.76$, $p = .016$ (two-tailed), since in the first pair infants looked equally long at the two test events, $p = .29$.

Non-parametric Wilcoxon signed ranks tests confirmed the results of the leader condition ($Z = 2.74$, $p = .006$) and bully condition ($Z = .98$, $p = .37$). Moreover, performing an ANCOVA using as covariate the infants' looking times during the first two familiarization trials revealed again a significant Trial × Condition interaction, $F(1, 28) = 7.19$, $p = .012$.

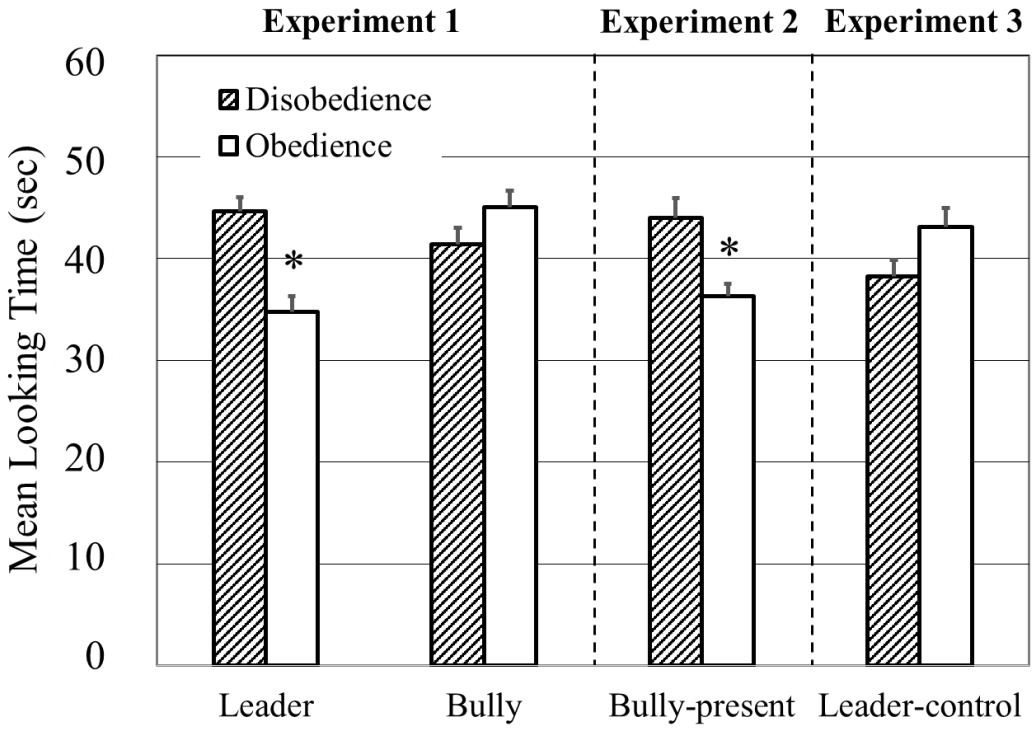

Figure 3. Results from Experiment 1-3. Mean looking time of the test events as a function of condition (Leader, Bully, Bully-present or Leader-control) and trial (Disobedience or Obedience). Error bars show the magnitude of the standard errors. * denotes a significant difference between trials within a condition, $p < .05$.

Discussion.

In the leader condition, infants looked reliably longer when subordinates did not comply with the instruction than when subordinates obeyed. This result suggests that infants expect subordinates to obey to the leader also in her absence. In the bully condition, infants looked equally at the obedience and disobedience events. This null result suggests that infants have no expectations when a dominant that use force to be followed commands to the subordinates and leaves them alone. Together, the results of Experiment 1 suggest that 21-month-olds distinguish from a simple form of dominance based of force (social dominance) and a more complex form of dominance (leadership), and predict that subordinates will obey a leader but not a bully in the absence of the dominant individual.

However, dominant individuals, which use force or intimidation, often succeed in constraining others to comply with their instructions. In Experiment 2, we then asked whether infants expect that subordinates will obey a bully that does not leave the scene. We predicted that infants would expect subordinates to comply (possibly because of the fear to be beaten again).

Experiment 2

Design. We asked whether 21-month-olds expect subordinates to comply with an order given by a bully in her presence. Infants watched identical computer animations to those used in bully condition of Experiment 1, except that the bully did not leave in test events. Test events (obedience, disobedience) order was counterbalanced between participants.

Method.

Participants. Participants were 16 healthy full-term infants, 7 male (20 months, 1 day to 24 months, 1 day, *M* = 21 months, 19 days). An additional 6 infants were excluded due to fussiness (3) or because they looked at the test events for the maximum time allowed (3).

Materials and procedure. The materials and procedure were identical to those used in Experiment 1. Interobserver agreement averaged 95% per trial per infants, and observers guessed the right order at chance level (.44). No significant interactions between trial and infants' sex or number of siblings were found, all *F*s < 1.

Results.

As predicted, infants looked reliable longer at the disobedience event (*M* = 44.1, *SD* = 15.7) than at the obedience event (*M* = 36.3, *SD* = 9.6), $t(15) = 2.71$, $p = .016$, $d = .72$ (two-tailed). As in leader condition of Experiment 1, the result was due to the difference between looking time in the second pair of test events, $t(13) = 3.93$, $p = .002$, since no significant difference was found in the first pair of test events, $p = .95$. A non-parametric Wilcoxon signed ranks test confirmed the result, $Z = 2.17$, $p = .03$.

Discussion.

In Experiment 2, we found that infants looked reliably longer when subordinates disobey to the bully's order in her presence than when they obey. Infants expected that subordinates would comply with the bully's instructions if she remained into the scene to watch and control them. Together with the results from Experiment 1, these findings suggest that 21-month-olds expect that subordinates will obey to the leader, also in her absence, but expect that subordinates will comply with a bully's instructions only in the case in which she is present and watching.

However, an alternative hypothesis could be put forward to explain our results from Experiment 1. Leader familiarization events involved a general positive interaction between

figures followed by a giving action from subordinates, but Bully familiarization involved a negative interaction followed by a bully's taking action (see Tatone, Geraci, & Csibra, 2015). Therefore, infants' expectations of obedience in Leader condition could have been formed because of the positive interaction between characters and the giving action and not because of the dominant agent being characterized as a leader. Experiment 3 aimed to address this criticism by asking whether infants expect that subordinates will obey a friendly figure that receive the ball but however is not a leader. We predicted that infants either would hold no expectations or would expect disobedience.

Experiment 3

Design. We asked whether 21-month-old infants expect characters to comply with an order given by a friendly agent that however was not characterized as a leader. Infants saw identical computer animations to those used in bully condition of Experiment 1, except that, in the character-familiarization event, the main agent came into the scene and said "hi" by waving a little, and the three red ovals responded by saying "hi" waving in turn; subordinates still gave the main character the ball and she left. Test events order was counterbalanced between participants.

Method.

Participants. Participants were 16 healthy full-term infants, 6 male (20 months to 23 months, 5 days, M = 20 months, 28 days). An additional 4 infants were excluded due to fussiness (2) or because they looked at the test events for the maximum time allowed (2).

Materials and procedure. Materials and procedure were identical to those used in Experiment 1. Interobserver agreement averaged 96% per trial per infants, and observers guessed the right order at chance level (.47). No significant interactions between trial and infants' sex or number of siblings were found, all Fs < 2.39.

Results.

Infants tended to look longer at the obedience event (M = 43.1, SD = 15.6) than at the disobedience event (M = 38.3, SD = 12.8), $t(15) = 2.02$, $p = .061$, $d = .54$ (two-tailed). The result was due to the significant difference between looking time in the second pair of test events, $t(14) = 2.33$, $p = .035$, since no significant difference was found in the first pair, $p = .87$. A non-parametric Wilcoxon signed ranks test confirmed the result, $Z = 1.89$, $p = .059$.

Discussion.

In Experiment 3, we found that infants tended to look longer when subordinates obey to the main character's instructions. Infants tended to expect that subordinates would not comply with the instructions of a main character who was not a leader although she was friendly and received the ball from the subordinates.

Previous studies on infant cognition show that early in life humans are able to represent positive and negative social interactions (e.g., Choi & Luo, 2015) and they interpret giving but not taking actions as inherently social (Tatone et al., 2015). Because Leader condition familiarization events in Experiment 1 differed from Bully condition familiarization events by involving a positive social interaction and giving actions, it was possible to hypothesize that infants' expectations were indeed a product of these different features of the events rather than being related to the leadership. However, in Experiment 3 we did not find that infants expected subordinates to comply with the main character's instructions. Therefore, a positive social interaction between figures and giving actions are not sufficient to generate infants' expectations of obedience.

General discussion

Across three experiments, we provided the first evidence suggesting that infants in the second year of life are able to distinguish leadership from a different type of social dominance based on force. We found that infants expect subordinates to comply with a leader's instruction, that is, they expect subordinates to obey to someone they bowed down to

and who bowed down in response (leader condition of Experiment 1). The null result in the bully condition of Experiment 1 helps us constraining our interpretation by showing that infants do not expect subordinates to comply with anyone's directives. Indeed, infants hold no expectations when the bully commanded to go to bed (bully condition of Experiment 1).

Moreover, we assessed that an exception is made when the bully remained to watch the subordinates complying with her instructions. Infants now expect again subordinates to comply, possibly because they expected that the subordinates would act out of fear of the bully (Experiment 2). Lastly, with Experiment 3 we rejected an alternative hypothesis that explained infants' expectations of obedience in Experiment 1 relying on contextual factors rather than leadership. Together, these results provide evidence of an early-emerging capacity to distinguish between two different forms of dominance, that is, social dominance by brute force, or bullying, and leadership, and show this distinction in their expectations regarding the obedient behavior of the subordinates.

Infants' representation of dominance structures and their behavior.

Recently, a growing interest in how infants' represent social dominance has lead researchers to report evidence of a human early-emerging capacity to rely on both agents' relative size and numerical group size to predict the outcome of a dominance relation. Infants expect bigger individuals (Thomsen et al., 2011), and individuals associated with larger groups (Pun et al., 2016), to prevail. Moreover, infants in their second year of life understand dominance as a relation between at least two individuals and expect such relation to be stable across different situations (Mascaro & Csibra, 2012).

Mascaro & Csibra (2012) showed that infants infer social dominance by witnessing the outcomes of conflicts (i.e., subordinates' deference) over the possession of an object or the occupation of a place. In a sense, we did a step back, since in our study we asked whether infants expect that deference and compliance would actually follow from a relation that was

characterized either as physical coercive (bullying) or as respectful (leadership). Our findings are consistent with previous results in suggesting an early-emerging capacity to represent social dominance relations, but also add to previous studies by showing that infants can understand a complex form of dominance such as leadership, and that they do not expect subordinates' obedience to an absent bully. By contrast, they expect that subordinates would obey an absent leader.

Our results that 21-month-olds distinguished between a bullying strategy and a leadership-followers relation can also be related to the studies on infants and toddlers' actual interactions with peers. In the second year of life, children's conflicts are shown to be over tangible resources such as toys (Bronson, 1975; Hay, 1984; Shantz, 1987), but nonetheless these conflicts have a clear social nature and they are not simply instrumental (Caplan, Vespo, Pedersen, & Hay, 1991; Cummins, 2006; Eckerman, Davis, & Didow, 1989; Hay & Ross, 1982). At this age, children both cooperate during play and conflict over toys. However, from the first to the second year of life, a developmental decrease in using force during conflicts and an increase in resolving disputes in a prosocial way have been reported (Caplan et al., 1991; see also Holmberg, 1980; Sackin & Thelen, 1984). Among the changes that could explain the development in conflicts resolution and social coordination (e.g., see Diamond, 2013, pp. 141-142; Onishi & Baillargeon, 2005), we can include toddlers' increasing imitation of others' actions, as, for example, in the 'follow-the-leader' game, in which children imitate in turns others' new play actions (Eckerman et al., 1989).

Given the rich and complex social life of children in their second year of life, where children do not limit themselves to dominate with force one another, to find an early-developing set of mental mechanisms responsible of representing simple as well as complex forms of social dominance relations do not entirely surprise us. Moreover, these mechanisms and this understanding may also constitute a basis for later changes in social interactions

occurring during childhood. Preschoolers' and older children' social behavior in conflicting situations is characterized by multi-dimensionality, but a standard result is that during childhood children's behaviors become less aggressive and coercive and more 'artful' and prosocial – from bullies to leaders (Coie & Dodge, 1983; Hawley, 1999, 2002; Killen & Turiel, 1991; LaFraniere & Charlesworth 1983; Parten, 1933; Strayer & Strayer, 1976; Wright, Zakriski, & Fisher, 1996).

Could the early interactions with peers affect infants' expectations about people's deference to dominant individuals? To explore this possibility, we performed an ANOVA on infants' looking time of Experiment 1, with trial (disobedience, obedience) as a within-subjects factor, condition (leader, bully) and child's attendance of a daycare institution (yes, no) as between-subjects factors. We found that Trial × Daycare attendance interaction tended to reach statistical significance, $F(1, 27) = 3.84$, $p = .06$. Further comparisons revealed that infants attending a daycare institution looked longer at the obedience event ($M = 48.2$, $SD = 12.5$) than at the disobedience event ($M = 38.3$, $SD = 11.5$), $t(8) = 2.31$, $p = .05$ (two-tailed); by contrast, infants that did not attend a daycare institution looked equally long at the two events, $p = .55$. These results suggest that infants who had the possibility to interact with peers because of their daycare institution attendance were more likely to expect subordinates to disobey to an absent bully. Social experiences may then facilitate the understanding that people imposing themselves by physical coercion not only are usually not followed by subordinates, but subordinates may even decide to oppose to them when they are absent. By contrast, leaders are followed and, as a result, they are likely to generate and maintain a stable social dominance hierarchy within a group.

However, the extent to which social exposure contributes to the understanding of the difference between a leader and a bully remains unclear. Future research should be devoted to the investigation of the eventual relationship between infants' expectations or mental

representations and their actual behavior, both when they interact with peers and adults. That is, more research is needed to elucidate whether a complex understanding such as the one we found in this group of 21-month-olds is reflected by the child's social interactions or play with peers or is merely the basis of a later development of a set of behavioral tendencies and mental understanding.

Social dominance and leadership—phylogenic and ontogenetic aspects.

Adults often distinguish between different forms of social asymmetries, and they use different strategies to attain the desired social rank (Cheng et al., 2013). A more long-term unstable strategy is to constrain others to deference by using physical force and intimidation. However, our evolutionary history has brought to our current societies also by selecting forms of dominance that facilitate the imitation process and skills acquiring and, as a result, the cultural learning process. Moreover, leadership or prestige permitted to increase the group cohesion and the in-group collaboration to reach shared and fundamental goals such as defending the group, managing the resource distribution and ensure the respect of cultural norms (Fiske, 1992; Henrich & Gil-White, 2001; Van Vugt, 2006).

Given the fundamental role of leadership in generating and maintaining our entire social life, we asked whether very early in life humans already possess the ability to represent leader-followers relations as distinct from social dominance relations. In particular, we found that obedience in the absence of the dominant figure is expected to a leader's instructions but not to a physically dominant's instructions. These results suggest that the basis for understanding the complex dynamics of power and deference to authority are already in play from an early age, and these implicit representations will likely inform and shape our rational and mature thinking about the correct ways to manage the society organization and, perhaps, will inform adult behaviors.

Here we provided an evolutionary account of children's early ability to distinguish between leaders and bullies. However, we also note that it cannot be excluded that an interaction between innate tendencies and cultural influences occurs to determine the expectations revealed for the first time by the current study. In fact, children at the end of their second year of life already have had several opportunities to interact with peers, if they attend a daycare institution, or to interact with others. Moreover, they have had the opportunity to learn from the social exchanges they witnessed. It is therefore a likely possibility that alongside an innate capacity to distinguish different forms of dominance, cultural environment represents the spark without which no real social understanding is possible.

Our findings are certainly only a first step towards the study of infants' competencies in representing different kinds of social power interactions. Future studies should for example investigate whether leadership is represented by infants as a relation between at least two individuals or as a stable individual property such as height or skin color. A further question would be whether infants have similar expectations to those who we reported also in the case of a leader-follower dyadic interaction, which may resemble the mother-child relation. Furthermore, an entire research project could be develop in order to reveal the mutual relationship between infants' implicit naïve sociology reasoning and their emerging adherence to social norms and adults' instructions on how to behave and to resolve conflicts with others. Lastly, a host of studies could be conducted in order to investigate dimensions of power asymmetries different to giving and complying with orders; for example, we should study infants' expectations and deontic implicit reasoning about the subordinates' willingness to imitate leaders, the leaders' power to set new norms, or the dominants' capacity to grant or deny permission.

Conclusion.

For the first time, here we reported that 21-month-old infants hold different expectations on the deference to bullying dominants and proper leaders. While they expected subordinates to comply with the leader's orders, they did not expect subordinates to obey to a bully's instruction, unless the bully forced subordinates to obey with her intimidating presence. These results suggest that infants possess an incipient understanding of the dynamics of power and social asymmetries that guide them in successfully navigate the social world. While being a first step, this study promises to have important implications for both educational programs in early infancy and future scientific inquiry on the early-emerging representational mechanisms underlying humans' reasoning in the socio-moral and sociological domains.

PART 2

CHILDHOOD: INTENT-BASED MORAL REASONING

CHAPTER 2

Children's Intention-based Moral Judgments of Helping Agents

This chapter is based on the following original article:

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

Abstract

During preschool years, children's disapprovals of harming actions increasingly rely on intention rather than outcome. Here we studied for the first time whether a similar outcome-to-intent shift occurs in their judgments of helping actions. Children aged four-to-eight (N = 404) were asked to evaluate the goodness and the deserved reward of attempted and accidental help (Experiment 1), and the badness and the punishability of attempted and accidental harm (Experiment 2). We found an outcome-to-intent shift both in goodness and badness evaluations. In judging attempts, children's intent-based goodness develops prior to the intent-based badness judgment. Contrary to previous results, we did not find any evidence that the intent-based judgment of goodness or badness constrains the development of the deserved reward or punishment judgment. These findings challenge recent theoretical proposals concerning the conceptual change and cognitive architecture underlying the development of moral judgment.

Children' Intention-based Moral Judgments of Helping Agents

In judging the morality of an action, people typically consider both its underlying intention and its external consequences. We may follow an *intentionalist ethics* and focus primarily on the intention, rather than the consequences (Abelard, 1971; Kant, 1785/1959). Consequences can be caused by luck, and luck is not a moral factor (Nagel, 1979; Williams, 1981). Or we may adopt a consequentialist ethics, e.g. the *ethics of responsibility*, which focuses primarily on the consequences of the action (Weber, 1919/1994). Whether it is right to clone humans does not seem to depend on scientists' intentions, but rather on the foreseeable practical consequences. By claiming that actions have moral value only with respect to the consequences they bring about, consequentialists are opposed to deontologists.

A major concern for deontologists is that valuing only consequences will result in justifying awful actions because they will bring about a greater good for some people. Unlike consequentialists, deontologists claim that some choices or actions are morally forbidden no matter what the consequences of these choices or actions will be. Thus, the role of intention and consequences in judging other's actions is at the core of the main theories in moral philosophy. A growing body of evidence shows that people's moral judgment is typically based on intentions, but it also relies on outcomes, especially when it is concerned with whether and how much to punish in cases of culpability (e.g., Berg-Cross, 1975; Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007; Gino, Shu, & Bazerman, 2010; Killen & Smetana, 2008; Kohlberg, 1969; Piaget, 1932).

The emergence of an intent-based moral judgment during childhood has been a core aspect of developmental theories since Piaget's (1932) seminal work. Piaget presented children with stories involving two characters: one who acted in a good-intentioned way but caused serious material damage, and one who acted in a bad-intentioned way but caused less serious damage. Piaget then asked children which character was naughtier and should be

punished. He reported a developmental change between ages 6 and 10 from a propensity to offer evaluations based on outcome to a propensity to offer evaluations based on intention.

This outcome-to-intent developmental shift has generally been found in a rich set of subsequent studies showing that younger children's moral judgments are more heavily influenced by outcomes than are older children's moral judgments (Armsby, 1971; Baird & Astington, 2004; Costanzo, Coie, Grumet, & Farnill, 1973; Helwig, Hildebrandt, & Turiel, 1995; Imamoglu, 1975; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Moran & O'Brien, 1983; Nobes, Panagiotaki, & Pawson, 2009; Surber, 1977; Yuill, 1984; Yuill & Perner, 1988; Zelazo, Helwig & Lau, 1996; Wainryb, Brehl, & Matwin, 2005). Ambiguous cases such as failed attempts to harm and accidental harm, where intentions and outcomes lead to conflicting responses, were particularly useful in revealing the outcome-to-intent shift. Research using these cases showed either younger preschoolers relying mostly on outcome (Helwig et al., 1995) or equally on intention and outcome (Cushman, Sheketoff, Wharton, & Carey, 2013; Killen et al., 2011).

Nevertheless, after the methodological limitations of Piaget's initial studies were overcome (Farnill, 1974; King, 1971; Nelson, 1980; see Karniol, 1978, for review), it became also clear that even preschoolers can use intent information to evaluate moral agents and actions, although it remains true that older children show greater sensitivity to mental states. In fact, Piaget's original tasks were not always suitable for assessing the use of intent cues by younger children, since they sometimes confounded intention and outcome, the agents' intentions were not stated explicitly, and the relevant information was difficult to remember (Turiel, 1983). Moreover, during some interviews, Piaget focused on what children thought or expected an adult (i.e., the father, the mother, or the schoolteacher) would do, not on what the child herself would do (e.g., punish or not). These shortcomings lead Piaget to underestimate preschoolers' ability to rely on intention when producing a moral judgment.

According to a recent dual-process model, children' and adults' moral judgments are best accounted for by assuming two distinct underlying processes, rather than a developmental replacement of a fully outcome-based moral reasoning by a fully intent-based moral reasoning (Cushman, 2008; Cushman et al., 2013). The intent-based process relies on the assessment of agents' mental states and on the automatic assignment of negative values to harmful actions to evaluate agents' moral character; the outcome-based process analyzes actions' outcomes to assess agents' causal responsibility. While badness judgments[1] are generated mostly by the intent-based process, punishability judgments are generated by both the intent-based process and the outcome-based process. In fact, by asking participants to evaluate the wrongness and the punishability of attempted but failed or accidental harming actions, Cushman (2008) found that wrongness (or badness) judgments rely mostly on mental states information, and punishment judgments rely on both mental states and consequences factors.

Evidence for this dual-process model comes also from neuroimaging studies showing activation of brain regions associated with cognitive conflict and top-down control when individuals judge ambiguous cases of accidental harm compared to cases of intentional harm (Young et al., 2007). Moreover, developmental research showed that the intent-based process does not develop simultaneously for attribution of badness and punishability; rather, it was suggested that it is the emergence of an intent-based badness judgment that constraints the development of an intent-based punishability judgment (Cushman et al., 2013).

Judging Harming and Helping Agents

Moral competence encompasses the evaluation of what is morally bad and wrong as well as what is morally good and just. However, the vast majority of studies have focused

[1] Judgments of 'badness' are not necessarily always moral judgments in a strict sense, as a toothache can be seen as bad, but not *morally* bad. In the scenarios used in the studies discussed and presented here, however, the context makes it likely a moral interpretation of 'bad' and experimenters avoided to specify 'moral' in their test questions.

selectively on evaluations of moral violations, neglecting to investigate how people produce evaluations of actions that are usually morally approved, or even admired, and how moral approvals develop during childhood. A recent and clear example of this bias in adult literature is the claim that the fundamental template unifying moral judgment is interpersonal harm (Gray, Young, & Waytz, 2012). One reason for this neglect in the current literature might be that people are more likely to produce a moral judgment when facing moral violations, rather than praiseworthy behaviors (Rosmini, 1840/1989).

Moving from adult to developmental literature, we found that moral competence is often conceptualized as the capacity to recognize moral transgressions as some acts that are "wrong because they have intrinsic effects for others' right and welfare" (Smetana, 2006, p. 121). Social domain theory maintains that morality is about the respect of fairness (Turiel, 2014). This view implies that moral violations involve a victim and are not contingent on a specific group consensus or authority mandate (unlike the social-conventional violations). This conceptualization has oriented researchers towards a rich set of novel and important research goals and led to a widespread consensus within the field of developmental moral psychology (Killen & Smetana, 2015). However, by building on this rich body of research findings, an extensive work remains to be done in order to reach an understanding of the child's judgment of moral approvals of helping actions that would be comparable to our understanding of the child's moral disapprovals of harming actions.

Helping and harming behaviors are sometimes conceptualized as two sides of the same coin (McGinley & Carlo, 2007), but there are important differences between them. Positive duties or duties of commission, such as 'be benevolent' or 'be charitable', appear to be less narrow, strict, and rigorous than negative duties, or prohibitions, such as 'do not murder' (Kant, 1785/1959). While positive duties do not usually prescribe any particular action and do not specify how much we ought to do, negative duties have less leeway with respect to

their violation. The command ‘do not lie’ is more precise and restrictive than the command ‘tell the truth’, despite the fact that they appear to be logical opposites. In certain occasions, we are free not to tell the truth by omission and out of prudence, but, according to Kant, we are never allowed to lie. We are freer in the ways we can fulfill our positive duties than we are in violating the constraints of our negative duties. Children, by second grade, appear to reason consistently with this distinction. They judge that refraining from harming is obligatory, and morally required, whereas helping is discretionary, and thus morally laudable (Kahn, 1992). Turning to judicial systems, in many countries there is no general duty to come to rescue a person in need, such as a victim of a car accident (Rosenbaum, 2004).

Numerous studies found noteworthy asymmetries in tasks requiring the processing of helpful and harmful actions (e.g., Bostyn & Roets, 2016; Knobe, 2003; Young, Scholz & Saxe, 2011). For example, harmful side effects are judged as produced intentionally more often than are helpful side effects (Knobe, 2003; Leslie, Knobe, & Cohen, 2006; Pellizzoni, Siegal & Surian, 2009). Negative cues appear to have a greater weight than positive ones (Rozin & Royzman, 2001), even for young children and preverbal babies (see Vaish, Grossman, & Woodward, 2008, for a review), and negative outcomes are much stronger cues to agency than are positive outcomes (Morewedge, 2009). These studies suggest that judgments of praiseworthiness and blameworthiness might follow different developmental pathways. If children consider negative duties as more restrictive than positive duties (Kahn, 1992), then they should judge harming actions more punishable than they deemed helping actions praiseworthy. However, the extent to which preschoolers and older children focus on consequences in evaluating prosocial actions as opposed to harmful actions remains unclear.

Here, we investigated the outcome-to-intent shift in judgments of moral goodness and deserved reward for helping behaviors. Our method was modeled closely on Cushman et al. (2013) in order to investigate whether their claims about judgments of harming behaviors

could be generalized to the development of judgments of helping behaviors. In Cushman et al. (2013), children ages 4 to 8 were asked to evaluate both the badness and the punishability of agents who attempted but failed to harm, and agents who accidentally caused some harm. Failed attempt cases (intention, no outcome) and accidental cases (no intention, outcome) are the most useful cases to investigate the interactions between the intent-based and outcome-based processes. These cases are instances where a dual-process model of moral judgment predicts a conflict between the two routes posited, one that attributes value to intentions and the other that attributes value to causal responsibility for outcomes. These two different processes generate opposite evaluations when some harm is causally determined, but not intended by the agent, and when no harm occurs, but the agent intended to produce it. Cushman et al. (2013) reported four main results:

1. Four-year-olds' judgments of agents' badness and punishability were more based on action outcome compared to older children's judgments; by age 5, most judgments of badness relied on intention. This change was particularly evident in the evaluations of accidents; the condemnation of accidental harm decreased with age, suggesting an increased sensitivity to the absence of any negative intention.

2. By age 5, the criteria to assess badness and punishability started to dissociate; an agent causing accidental harm was judged more punishable than bad, suggesting that badness judgment became intent-based, while punishment judgment remained more outcome-based than badness judgment.

3. The intent-based attribution of badness mediated the effect of age on punishability: children first started to produce intent-based badness judgments and only later, as a consequence, produced intent-based punishment judgments.

4. In accidental harm scenarios, from age 5, badness judgments affected the subsequent punishability judgments, facilitating intent-based responses, but not vice versa (i.e., evaluating the punishability of an action had no effect on the subsequent badness attribution).

According to the authors, the evidence in (3) and (4) provides support for their developmental constraint model: intent-based badness judgments increasingly constrain punishability judgments to rely on intent as well. According to the authors, these results are due to a developmental change in the concept of what is morally bad rather than solely to changes in children's theory of mind (Chandler, Sokol, & Hallett, 2001; Killen et al., 2011; Smetana, Jambon, Conry-Murray, & Sturge-Apple, 2012) or executive function (Richardson, Mulvey, & Killen, 2012; Zelazo et al., 1996). This conceptual change involves a shift from a 'consequentialist' concept of wrong (an action is wrong if it causes negative outcomes) to a different, intent-based concept (an action is wrong if it was motivated by negative intentions).

Conversely, a continuity hypothesis posits that changes external to the moral domain explain the occurrence of the shift and that those changes reveal a latent conceptual repertoire and a capacity for intent-based moral judgment. An important set of changes may take place in executive functioning skills. Improvements in the ability to integrate different information, to select correct responses, and to inhibit wrong ones, could be fundamental in causing the shift (Margoni & Surian, 2016a). A second possible causal factor could be the development of theory of mind. Neuroimaging evidence in adults showed an association between moral judgment and theory of mind (Chakroff & Young, 2015). Blame of accidental harms is inversely associated with the activation of right temporoparietal junction (TPJ), a brain area selectively involved in mental state reasoning (Young & Saxe, 2009) and the inhibition of TPJ disrupts mental state reasoning during a moral judgment task (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). Further evidence came from studies of individuals with autism who did not judge accidental and attempted cases as morally different,

suggesting that they fail to take into account the mental state information (Moran, Young, Saxe, Lee, O'Young, Mavros, & Gabrieli, 2011). Moreover, since the theory of mind skills are linked with the development of executive functions, the two explanations may be functionally related (Leslie & Polizzi, 1998).

This Study

Do similar changes take place in judging helping and harming agents? The present research explored this question by examining children's evaluations of helping and harming agents. In Experiment 1, we presented children with scenarios involving failed attempts to help and accidental help. In Experiment 2, children were asked to judge examples of failed attempts to harm and accidental harm.

The primary aim of our study was to describe the occurrence of an outcome-to-intent shift in children's goodness judgment and to replicate previous results on judgments of harming actions. Such a shift should be revealed by an age-related increase in attributing goodness to cases of attempted but failed help and a decrease in attributing goodness to cases of accidental help. The differences between positive and negative duties discussed above motivate a detailed investigation of possible differences in the outcome-to-intent shift for goodness and badness attributions, in scenarios involving helping and harming actions, respectively. We wished to assess whether the intent-based goodness judgment develops approximately at the same time, before or later than the intent-based badness judgment.

The second aim concerns the explanation of the shift. Since we assessed the extent to which the pattern of results previously reported in the case of evaluation of harmful agents mirrors the results on children's evaluation of helping agents, here we provide further evidence to decide between the view positing a conceptual change (the '*constraint hypothesis*') and the view positing conceptual continuity (the '*continuity hypothesis*'). These

hypotheses posit different mechanisms explaining the development of the well-known connection between moral judgment and intentionality (Bloom & Wynn, 2016).

Extending the *constraint hypothesis* and the dual-process cognitive model to the evaluation of helping cases, if an outcome-to-intent shift occurs in the development of moral approvals similarly to moral disapprovals, the emergence of intent-based goodness judgment should constrain the emergence of intent-based deserved reward judgment, and the effect of age on reward judgments should be mediated by intent-based goodness judgment. This hypothesis maintains that theory of mind and executive functioning skills are prerequisite for the acquisition of a mature moral reasoning. However, it claims that the preschoolers' concept of badness (or goodness) is not the same concept older children have. The emerging intent-based concept would drive the development of punishability or deserved reward attributions. Younger children would judge both goodness and deserved reward relying on outcomes, but older children would develop an intent-based goodness judgment and, gradually, what deserves a reward would follow, in part, from what is good.

By contrast, the *continuity hypothesis* claims that the concept of badness (or goodness) remains the same throughout the development, and changes occurring outside the moral cognition allow the children to express a latent conceptual repertoire. If this is the case, following Cushman et al. (2013), one should expect similar, concurrent, but independent, developmental changes in goodness and deserved reward judgment, since domain general changes should affect similarly the two kinds of judgments. An explanation based on developing working memory, executive control and mental state reasoning skills would not predict any specific constraint of one type of judgment on the other judgment, but rather would be more consistent with a correlated and simultaneous shift of both kinds of judgments, resulting from domain general changes in the cognitive system.

Experiment 1

Method. Participants were 404 children ranging in age from four to eight years (age 4 *n* = 86; age 5 *n* = 84; age 6 *n* = 84; age 7 *n* = 78; age 8 *n* = 72), 187 female. The sample size was determined by running a-priori sample size calculation. Participants were recruited in several different nursery schools and elementary schools nearby Trento, in Italy. All children were Italian native speakers and no one was affected by sensory or cognitive impairments. Children attended schools serving a middle-income population and almost all of them were Caucasian. The parents gave their written informed consent. The Human Research Ethics Committee of the University of Trento approved the experimental procedure.

We also interviewed 15 four-year-olds and 19 five-year-olds in a pilot study. An additional adult sample (*n* = 24, all female) participated. Adults were students from an introductory psychology course at the University of Trento (mean age = 21.83 years, *SD* = 2.1), and all of them were Caucasian.

Materials and procedure: main study. The experimental sessions started with a 'warm-up' story involving a non-moral event in which a character wants to get a fish out of the aquarium tank but ends with getting another fish. This story had a similar structure to the following moral stories. The story was read aloud by the experimenter. Children were then asked two comprehension probes, one about the character's intention and the other about the outcome, presented in a counterbalanced order.

Children were then presented individually with two stories, one involving a failed attempt to help (i.e., good intention but no consequence), and one involving an accidental help (i.e., good consequence without relevant intention). We focused on helping actions because they are typically approved, and because help is conceptually the opposite of harm. Recent evidence suggests that even preverbal infants attribute a positive value to helping actions and a negative value to harming actions (Hamlin, Wynn, & Bloom, 2007; Hamlin & Wynn, 2011; Hamlin, 2013).

Stories were read aloud to the children by the experimenter and with the help of four vignettes illustrating the main phases of the stories. Pictures were used to reduce working memory demands, while the verbal text made relevant information explicit and salient. The vignettes were placed in front of the child, one at a time, at the appropriate point of the story as the experimenter read. The last vignette remained in front of the child when children were asked the test questions. Stories were drawn from two story contexts. Each context had two versions: a failed attempted helping action or an accidental helping action. Below we report the synopses of the stories used.

'Tree stories'. Attempted help: one boy wants to retrieve his little brother's lost ball that is on a tree by hitting it with his favorite ball, but the boy sneezes while trying to retrieve the ball and fails. Accidental help: one boy is playing with his favorite ball when he sneezes and accidentally hits and retrieves the lost ball of his brother.

'Door stories'. Attempted help: one boy wants to open a heavy door for his little brother, but the boy stumbles and fails. Accidental help: one boy is running around the room when he stumbles and accidentally opens a heavy door for his little brother.

Each child received one 'tree' and one 'door story', which varied in terms of whether they describe an accidental or a failed attempted helping action. After each story, children were asked two comprehension questions whose aim was also to drive children's attention on features relevant to our study. One question was on whether the character wanted to produce the outcome, and the other one was on whether he actually produced the outcome. To children who failed at least one of the probes, the experimenter offered the opportunity to listen to the story again, and the same probe questions were asked. Subsequently, children were asked two test questions: "According to you, in the story I just told you, is [character name] a good boy or not?", and "According to you, in the story I just told you, does [character name] deserve a reward or to be thanked, or not?" After both stories, children were

asked to say which character they think was ‘più buono’ (literally, ‘more good’): “According to you, in the two stories I told you, who is ‘more good’ between [the characters]?” The order of presentation of stories (attempted help vs. accidental help story), comprehension probes, and test questions (goodness vs. deserved reward) were counterbalanced.

Stimuli and procedure were modeled on Cushman et al. (2013). However, we introduced some changes. We stated explicitly how the character who was helped (or not helped) felt in response to the event. We did so to improve the comprehension of the stories (Grueneich, 1982; Stein & Glenn, 1979), and to illustrate how the consequences were related to agents’ wellbeing (Arsenio, 1988; Smetana, 2006). To gain further insight on the weight the child attributes to intentions and outcomes, we added a choice task by asking which of the two characters was ‘more good’. To avoid problems related to judging two very similar stories, with the risk of unwanted comparisons along dimensions that are irrelevant for the aim of the present study, we read two stories drawn from two different contexts.

Preliminary study. To check whether younger children understand the main elements of the stories, we administered to 15 four-year-olds and 19 five-year-olds two morally unambiguous versions of the ‘tree story’. One story involved an event of attempted and succeeded help (one boy wants to retrieve the ball of his brother and he succeeds). Another story involved an event in which the character has no intention to help and does not help (one boy does not see his brother and just wants to play with his ball; meanwhile he is playing, his brother tries to retrieve the lost ball). The procedure remained the one followed in the main study. Order of stories presentation and order of comprehension probes were counterbalanced.

Adult study. Adults’ judgments were assessed with the same stories and questions used with children. The students were tested in small groups and they were asked to read the stories and express their evaluations on a sheet. The students were informed that the materials

were used in a study on children, in order to justify the childish aspects of the stories. Stories and questions were presented in a counterbalanced order.

Results.

We present the results following closely Cushman et al. (2013)'s article that inspired the present research. In a first part, we focus on the analyses to test the occurrence of an outcome-to-intent shift in moral reasoning. In a second part, we focus on the analyses relevant to test the constraint hypothesis. In the results sections of Experiment 1 and 2, when performing multiple comparisons, we adjusted the alpha level applying the Bonferroni correction.

Preliminary study. In a preliminary study, we tested children's comprehension of the main task by asking to a separate group of younger children (aged 4 and 5) to evaluate two morally unambiguous stories. One 4-year-old failed one of the two control questions of one story and no child failed both. Children always judged the character who intentionally helped to be good and almost never judged the character who did not help to be good (97% and 9%, respectively), McNemar $\chi^2(1, N = 31) = 25.04$, $p < .001$. In addition, they judged that the character who intentionally helped was worthy of a reward (94%), but the character who did not help was not (6%), McNemar $\chi^2(1, N = 31) = 23.31$, $p < .001$.

Main study. We excluded the responses to the stories for which a child did not pass both control questions (age 4 = 15% of responses; age 5 = 2%).

Preliminary analyses on story order effects.

Before proceeding, we tested whether children's responses were affected by story order, conducting two factorial logistic regression analyses that showed that the effect of age on approval judgments of accidental help or attempted help does not interact with the order in which the stories were presented (all *p*s for interactions > .11). Since children's approvals were not affected by story order, we will consider children's responses to both stories for the

following analyses. The term ‘approval judgment’ refers to a response obtained by collapsing between goodness and deserved reward judgments: the child approved the character when she attributed goodness or deserved reward, and did not approve when she did not attribute neither goodness nor deserved reward. The two judgments were collapsed to facilitate the comparison with previous work in which badness and punishability judgments were initially collapsed in a single dependent measure, and story order effects were tested using that measure.

Adults’ approvals of accidental help were instead affected by the order of story presentation; 67% of adults judged the accidental help approvable in the case they had to judge it first, but only 8% judged it approvable in the case they had to judge it second, following the attempted help, $\chi^2(1, N = 24) = 8.71$, $p = .003$.

Testing the occurrence of an outcome-to-intent shift.

We focused our analyses on each child’s first response in order to eliminate test questions order effects (see also 2.2.2.3. session below).

The choice measure (‘more good question’). A first piece of evidence for an outcome-to-intent developmental shift in goodness attribution can be easily observed on Figure 1a, reporting children’s answers to the question about which of the two stories characters is ‘more good’. A logistic regression analysis on children’s responses, with age group as predictor, revealed an age-related increase in pointing to the character who attempted to help, $\beta = .65$, 95% CI [.47, .84], $z = 6.81$, $p < .001$.

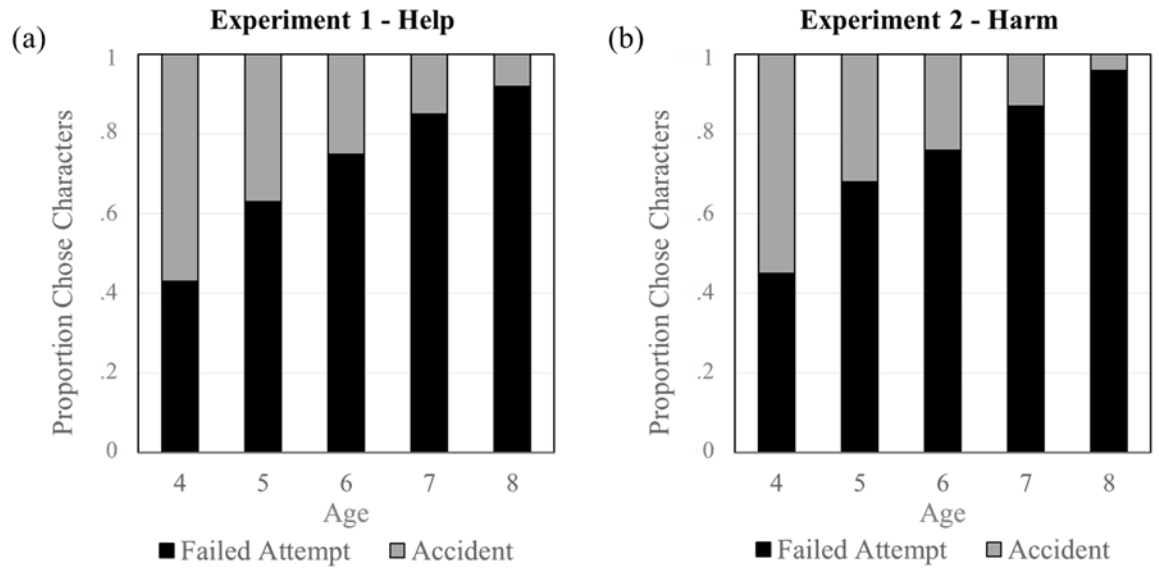

Figure 1. (a) Proportion of children who chose the character who attempted, but failed to help or the character who accidentally helped, when asked to choose which one was 'more good' (Experiment 1). (b) Proportion of children who chose the character who attempted, but failed to harm or the character who accidentally harmed, when asked to choose which one was 'more bad' (Experiment 2).

Development of goodness and reward judgments. Children's first responses on both types of stories are illustrated in Figure 2a. Performing a series of logistic regression analyses on both goodness and deserved reward judgments, using age group as predictor, we found an age-related increase in goodness attribution to attempted help, β = .59, 95% CI [.17, 1.01], z = 2.76, p = .006, and a decrease in goodness attribution to accidental help, β = .36, 95% CI [.12, .60], z = 2.89, p = .004. By contrast, age-related changes in deserved reward judgments were not significant. We further noticed that goodness judgment did not decrease between 4 and 6 years (p = .229), and 4-year-olds attributed goodness to the character who attempted to help more often than expected by chance (74%, p from binomial test = .005), but they did not attribute goodness to the accidental story character less often than expected by chance (43%).

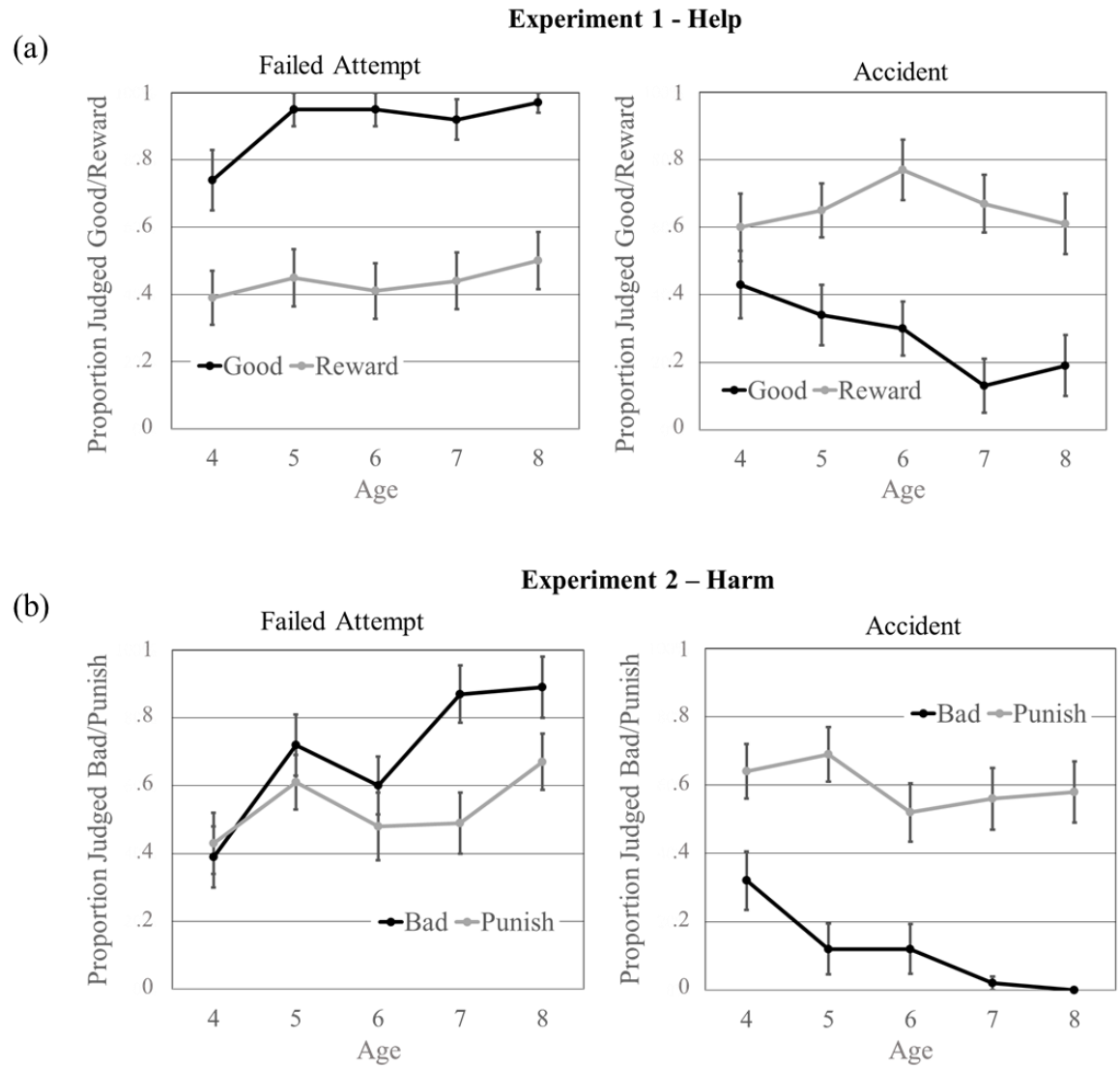

Figure 2. Proportion of participants who judged the character good or worthy of a reward (a), or bad or punishable (b), including data from each child's first response. Error bars show the magnitude of the standard error.

Attempted vs. accidental help. In each age group, children judged attempted help to be more good than accidental help, age 4: *p* from binomial test = .003; age 5: McNemar $\chi^2(1, N = 41) = 21.33$; age 6: McNemar $\chi^2(1, N = 40) = 24.04$; age 7: McNemar $\chi^2(1, N = 39) = 27.27$; age 8: McNemar $\chi^2(1, N = 36) = 26.04$; all *p*s < .003. This difference increased with age, from 31% of difference at 4 to 78% at 8 years (see Table 1 showing proportion of 'yes'

answers to test questions, including data from each child's first response). A regression analysis examined the effect of age on the difference scores between goodness judgments of attempts and goodness judgments of accidents, revealing a significant effect of age, $\beta = .11$, 95% CI [.06, .16], $z = 4.23$, $p < .001$. Conversely, in each age group, children did not judge attempted help to be less worthy of a reward than accidental help. Adults' attributions of goodness in attempted help stories were significantly more frequent than in accidental help stories (100% vs. 17%, p from binomial test = .002), whereas attributions of deserved reward did not differ significantly (67% vs. 33%, $p = .289$).

Table 1

Proportion of children who judged the character good or worthy of a reward (Experiment 1), or bad and punishable (Experiment 2).

	Experiment 1 (Helping Agent)				**Experiment 2 (Harming Agent)**			
	Goodness		Reward		Badness		Punishment	
Age (years)	Attempted	Accidental	Attempted	Accidental	Attempted	Accidental	Attempted	Accidental
4	.74	.43	.39	.60	.39	.32	.43	.64
5	.95	.33	.45	.65	.72	.12	.61	.69
6	.95	.30	.41	.77	.60	.12	.48	.52
7	.92	.13	.44	.67	.87	.3	.49	.56
8	.97	.19	.50	.61	.89	.0	.67	.58

Attributions of goodness vs. deserved reward. If evaluations of helping actions follow the same pattern reported previously for harming actions, then the criteria to assess goodness and deserved reward should start to dissociate roughly by age 5. We should find that attempted help (good intention) is judged more good than worthy of a reward, since goodness judgments become predominantly intent-based, and accidental help (good outcome) is judged

more worthy of a reward than good, since reward judgments rely both on intent and outcome. We found the predicted dissociation in attempted help by age 4, and in accidental help by age 5 (see Figure 2a). In *attempted help*, goodness attributions were significantly more frequent than deserved reward attributions in each age group, age 4: $\chi^2(1, N = 74) = 9.12$; age 5: $\chi^2(1, N = 84) = 25.12$; age 6: $\chi^2(1, N = 84) = 27.59$; age 7: $\chi^2(1, N = 78) = 21.25$; age 8: $\chi^2(1, N = 72) = 20.66$; all *p*s < .003. In *accidental help*, by 5 years children's goodness attributions were less frequent than reward attributions, age 5: $\chi^2(1, N = 82) = 8.22$; age 6: $\chi^2(1, N = 84) = 18.90$; age 7: $\chi^2(1, N = 78) = 23.61$; age 8: $\chi^2(1, N = 72) = 12.99$; all *p*s < .004. Adults judged attempted help more good than worthy of a reward, $\chi^2(1, N = 24) = 4.8$, $p = .028$, whereas they judged accidental help as good as worthy of a reward.

Testing the constraint hypothesis.

According to Cushman et al. (2013), the continuity hypothesis predicts that a) test questions order should have similar effects of one type of judgment (goodness or reward) on the other type of judgment and that b) the age effect on one type of judgment should not be mediated by the development of the other type of judgment. In other words, the continuity hypothesis runs against the expectation that children's responses will become more intent-based on one type of judgment and, as a result of that, will become more intent-based on the other type of judgment.

Finding that test questions order effects are significant, regardless of which test question is asked first, and that the effect of age on reward judgments is not mediated by intent-based goodness judgment, would run against the constraint hypothesis, which predicts the opposite results. Such pattern of findings would be more consistent with an alternative view based on conceptual continuity, but it will not be supporting, in the absence of any independent measure of executive functions or theory of mind skills, any particular version of it.

Analysis of mediation effects. We tested whether the intent-based goodness judgment mediated the effect of age on deserved reward judgment, as predicted by the constraint hypothesis. In children's responses, we found only two significant correlations, both between age and goodness judgment, one in the attempted help stories $r = .21$, $p < .001$, and the other in the accidental help stories $r = -.29$, $p < .001$. Since we did not find any correlation between age and deserved reward, we failed to find any support for the constraint hypothesis, which predicts that the effect of age on deserved reward attributions is mediated by the intent-based goodness attributions.

Test question order effects. Competing accounts of the developmental shift were further tested by examining the effect of test questions order on attributions of goodness and deserved reward. While the continuity hypothesis is consistent with a bidirectional influence between judgments of goodness and deserved reward, the constraint hypothesis is consistent with a unidirectional influence between judgments. We report only the effects on accidental help evaluations, since we found no test questions order effect on attempted help evaluations. The analyses we report compare three age groups (4, 5 and 6-8 years old), to facilitate the comparison with previous results from Cushman et al. (2013).

In younger children, the order of questions affected goodness judgments of accidental help, but not deserved reward judgments (Figure 3a). Four- and five-year-olds who made judgment of deserved reward first were more likely to judge the character who helped accidentally to be good compared with those who made goodness judgment first; 4 years: 71% vs. 43%, $\chi^2(1, N = 72) = 5.83$, $p = .016$; 5 years: 68% vs. 34%, $\chi^2(1, N = 81) = 9.01$, $p = .003$. By contrast, the order of questions did not affect younger children's judgments of deserved reward.

In older children, the order of questions affected both goodness and deserved reward judgments. Deserved reward judgments affected goodness judgments, $\chi^2(1, N = 234) = 8.07$,

p = .004, and goodness judgments affected deserved reward judgments, $\chi^2(1, N = 234) = 9.88$, p = .002. That is, children who made deserved reward judgment first were more likely to judge the character to be good compared with those who made goodness judgment first, and children who made judgment of goodness first were less likely to judge the character to be worthy of reward compared with those who made deserved reward judgment first. The order of questions did not affect adults' judgments.

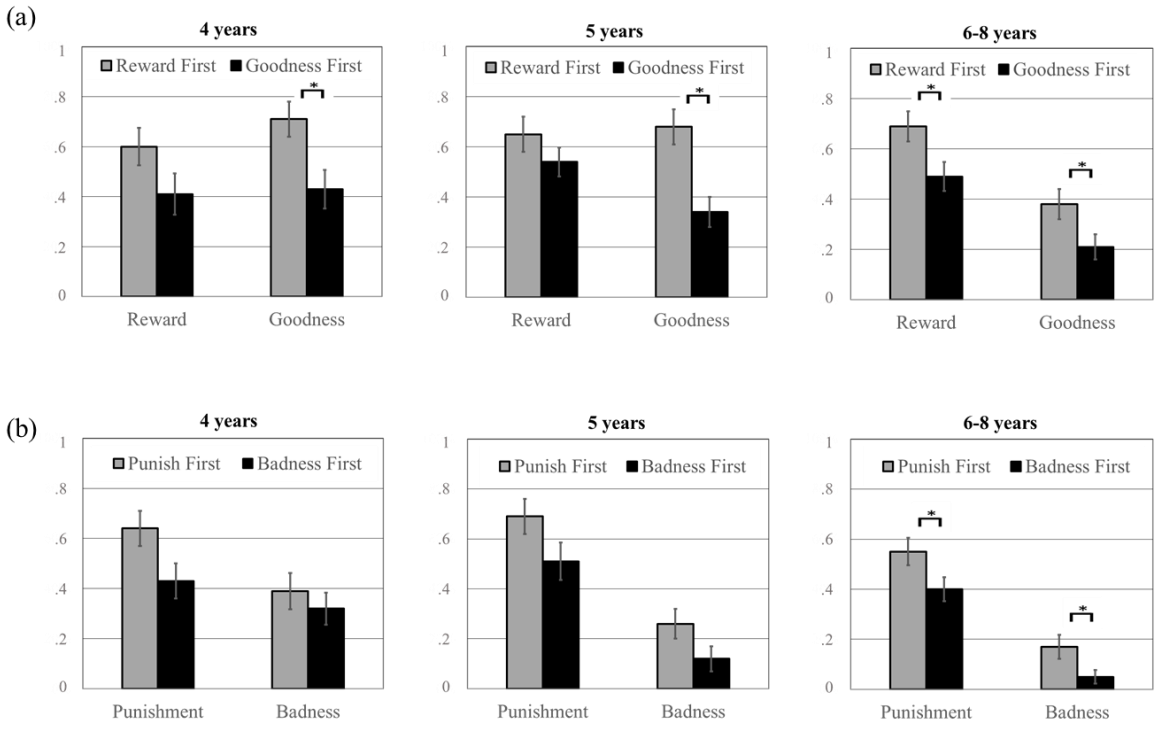

Figure 3. *Accidental help and harm trials.* (a) Proportion of participants who judged the character to be good and to be worthy of a reward at 4, 5, and 6-8 years as a function of the questions order (Experiment 1). (b) Proportion of participants who judged the character to be bad and to be punishable at 4, 5, and 6-8 years as a function of the questions order (Experiment 2). Error bars show the standard error. * $p < .05$.

Discussion.

We found evidence of an outcome-to-intent shift in goodness judgments of both failed attempts to help and accidental helps. First, when asked to choose which character was 'more good', older children pointed to the character who attempted to help (good intention, no outcome) more often than younger children did, revealing a growing sensitivity and reliance on intention. Second, goodness attribution to attempted helps increased with age, mostly due to the developmental change occurring between the age of 4 and 5. Third, goodness attribution to accidental helps decreased, showing a developmental change occurring at 6-8 years (see Figure 2a).

Younger children's sensitivity to intention.

With increasing age, moral goodness judgment relies more on intention and less on outcome. Note that even 4-year-olds judged attempted help (good intention) more good than accidental help (good outcome). This difference increased with age, revealing a growing intent-based judgment. While this is evidence of an outcome-to-intent shift, it does not imply that younger children do not take into account mental state information in their judgments. In fact, one can also easily notice that in judging attempted help, 74% of the 4-year-olds attributed goodness to the story character. This is consistent with previous research suggesting that even younger preschoolers are sensitive and can use intention information in their moral evaluations when adequately assessed (Imamoglu, 1975; Karniol, 1978; Nelson, 1980; see also Armsby, 1971; Farnill, 1974; Nobes et al. 2009; Yuill & Perner, 1988).

In younger children, the sensitivity to intention was found in evaluating attempted help, i.e., where the intention cue was particularly salient, and was not found in evaluating accidental help, where the good outcome was presumably a more salient cue than the agent's intention. The likelihood to attribute goodness to the character who attempted to help increased during preschool years, and was already above chance level at 4 years, but the

likelihood to attribute goodness to the character who helped accidentally decreased during primary school years. Until 5-6 years, the outcome information maintains some relevant influence on moral reasoning, when children are asked to evaluate accidental helping actions. Consistently, while in the case of accidental help, 4-year-olds did not yet judged the character less good than worthy of a reward, in the case of attempted help even 4-year-olds judged the character more good than worthy of a reward.

Differences with previous results on the evaluation of harming actions.

Although our main results concerning the description of the outcome-to-intent shift mirror, on several respects, previous results on harming actions evaluations, the pattern of results from Experiment 1 differs from Cushman et al. (2013)'s in a number of interesting ways. While we found both an age-related increase in goodness attribution to attempted help and a decrease in goodness attribution to accidental help, they only reported a decrease in badness and punishability attributions to accidental harm. In the responses to accidents, similarly to the case of harming evaluations, we found that goodness attributions start to decrease at 5-6 years. The criteria to attribute goodness and deserved reward to accidents start to dissociate at 5 years, that is, the same age when badness and punishability star to dissociate. Moreover, in line with previous results on punishability and badness attributions, showing attribution of badness to be mainly intent-based and attribution of punishability to be both intent- and outcome-based (Cushman, 2008; Cushman et al., 2013), here we found that children's deserved reward judgment was more outcome-based than goodness judgment. However, unlike previous studies showing that punishability judgments change with age, we found that deserved reward judgment does not change with age. In sum, we found two main differences concerning the outcome-to-intent shift between our study and the previous study on harm evaluations: a) a developmental change was found in both evaluations of accidental and attempted help; b) we found no change in deserved reward attributions.

Before discussing the implications of these differences with previous results (see discussion session of Exp. 2), we want to assess whether we find a similar pattern of results for punishability judgments (that is, no developmental changes in punishability judgments) by using the same procedure we used in Experiment 1. If this will be the case, then it will be easier to interpret the differences between our findings and previous work also by pointing out some relevant procedural aspects (see section 3.3.2).

Table 2

A comparison of the main results found in the present study and by Cushman et al. (2013).

	Outcome-to-intent shift		Constraint hypothesis	
	1.	2.	3.	4.
Experiment	Decreasing disapprovals of accidents and increasing disapprovals of attempts	Increasing dissociation between badness and punishability judgments	Intent-based badness judgment mediates the relationship between age and punishability	After the shift occurred, intent-based badness judgments constrain punishability judgments, but not vice versa
Cushman et al. (2013)	✓	✓	✓	✓
Present study (Exp. 2)*	✓	✓	×	×

Note. ✓, the effect was found; ×, the effect was not found.
*In the present study, the pattern of results supporting the outcome-to-intent shift but not the constraint hypothesis was also found in approval judgments (Exp. 1).

Questions order and mediation effects.

In analyzing our data, we focused on each child's first responses. This was done to eliminate possible order effects. In fact, beginning with younger children's results, we showed that at 4-5 years, deserved reward judgment constrained the subsequent goodness judgment of accidental helping action, and children appeared to endorse the 'realist' criterion "the agent deserved a reward, therefore she must be good" (Kohlberg, 1969; Piaget, 1932).

While deserved reward attribution constrained goodness attribution, the reverse was not true. This pattern is consistent with the tendency of younger children to rely also on outcome, particularly when they evaluate the morality of accidental helping actions. Older children's judgments of accidental help were affected also by the opposite order effect (i.e., goodness judgments affected deserved reward judgments), suggesting a growing flexibility and context-sensitivity in judging moral scenarios (Turiel, 1983; Jambon & Smetana, 2013). Further evidence for context-sensitivity was found analyzing stories order effects on adults' responses. Adults who judged attempted help first were much less likely to approve accidental help compared with those adults who judged accidental help first, probably because they focused on the absence of any good intention, which was highlighted by the previous judgment of attempted help.

Alongside these order effects results, we also failed to find that the intent-based goodness judgment mediated the effect of age on reward judgments, since we did not find any age-related change regarding deserved reward attributions.

Do the results from Experiment 1 support the constraint hypothesis and a view based on conceptual changes? The core prediction of the constraint model of moral judgment development (Cushman et al., 2013) is that the new emerging intent-based badness judgment should constrain the punishability judgment, but not vice versa. Consistently, analyzing questions order effects, one should find that the intent-based judgment of goodness (here) would affect the subsequent judgment of deserved reward, but not vice versa.

The results of Experiment 1 do not support the constraint hypothesis, because a) we did not find the mediation effects predicted by such hypothesis, and b) in older children, once the shift occurred (i.e., when the model predicts that a new concept of goodness should be developed), we found a bidirectional influence between judgments of goodness and deserved reward, where the constraint hypothesis predicts instead a unidirectional influence. Then,

these results appear to be more consistent with the alternative hypothesis based on conceptual continuity (e.g., Chandler et al., 2001; Killen et al., 2011; Richardson et al., 2012; Smetana et al., 2012). In fact, no constraining effect or, if any, a bidirectional constraining effect between judgments of goodness and judgments of deserved reward are predicted by the hypothesis that ancillary changes occurring outside the moral domain do not affect selectively the moral concept of goodness or deserved reward. A bidirectional constraining effect may indicate that there is not a judgment that prevails selectively over the other and determines its development, and at the same time may simply indicate that our task was more permeable to priming effects. The evidence from Experiment 1 runs against the constraint hypothesis, but does not point out what factors, in a continuity view, are responsible for the observed changes. Future studies should investigate directly the continuity hypothesis by testing which changes in social and cognitive abilities are associated with changes in moral judgment of helping agents.

The constraint hypothesis predicts that, after the shift has occurred, the intent-based goodness judgment would constrain the deserved reward judgment to be also intent-based. Then, one should also expect that such a constraining effect is not present in younger children's responses, *before* the occurrence of the shift. In this respect, the unidirectional order effect we found in younger children may still be consistent with the constraint hypothesis, since it was not the goodness judgment to influence the subsequent reward judgment. Finding that deserved reward judgment constrained goodness judgment can indeed be accommodated by a theory of development that poses two independent cognitive processes (intent- and outcome-based), and that predicts that younger children's responses rely also on outcome. If younger children's judgments are more outcome-based than older children's judgments, then it is not surprising to find that outcome-based deserved reward judgments constrain goodness judgments. However, finding that older children's judgments

of goodness and deserved reward constrained each other, along with finding no mediation effects, provides evidence against the idea that the acquisition of an intent-based goodness judgment selectively constrains the emergence of an intent-based reward judgment. Then, the comparison of the results of Experiment 1 with the results from Cushman et al. (2013) might suggest that judgments of positive and negative cases follow different developmental pathways.

One result in the present study appears at odds with the claim that the shift is due to changes external to the moral domain: we found developmental change in goodness attributions, but not in deserved reward attributions. This result may be seen as consistent with a shift that reflects a selective change in the concept of goodness, that is, from an initial outcome-based concept of morally good to a later intent-based concept. One may ask why changes in domain general executive functioning skills, or in theory of mind, would bring about a developmental change in goodness attributions, but not in deserved reward attributions. However, we doubt that the answer to this question will provide support for the constraint model of moral judgment development. In fact, this model, in the case of approvals, does not simply predict a selective change in goodness judgment, but predicts that goodness judgment becomes intent-based *before* deserved reward judgment and constrains the latter to become also intent-based. Since we found that the frequency of children's intent-based deserved reward judgments never changed with age and that reward judgments constrained goodness judgments (in older children), the constraint model needs to be revised in order to account for the present findings. In sum, this pattern of findings suggests that a simple generalization of the view put forward by Cushman et al. (2013) to the positive cases does not work.

Conclusions.

In Experiment 1, we reported evidence of an outcome-to-intent shift in goodness judgments of helping actions. Moreover, we suggest that processing changes occurring outside the moral domain, such as in theory of mind and executive control, not only help to explain the outcome-to-intent shift in approval judgments, but an explanation based on these processing factors can accommodate a large part of our results. These results do not mirror the pattern of previous results on children's disapprovals that supported the constraint hypothesis (from Cushman et al., 2013). However, because we introduced some procedural changes with respect to such previous work on disapprovals, it is desirable to provide a more stringent comparison between positive and negative cases evaluations by employing the same procedure.

Experiment 2

In Experiment 2, the same procedure followed in Experiment 1 was used to elicit evaluations of harming actions.

Method. All children who participated in Experiment 1, participated in Experiment 2 (N = 404); 214 children were firstly interviewed about the goodness and the deserved reward of helping behaviors, and roughly after three weeks were interviewed about the badness and the punishability of harming behaviors, while 190 were interviewed in reverse order.

Materials and procedure. We used materials and procedure very similar to Cushman et al. (2013), but the following changes were introduced. Stories of attempted and accidental harm were read to the children with the help of four vignettes, and not three as in the previous study. We used a fourth picture because we added to the stories a description of how the harmed (or not harmed) character felt in response to the event. We did so to help younger children to comprehend stories in their moral relevance (Arsenio, 1988; Smetana, 2006). We used two story contexts from Cushman et al. (2013).

'Push stories' - Attempted help: one boy attempts to push somebody over when he trips on a rock and misses. Accidental help: one boy is running when he trips on a rock and accidentally pushes somebody over.

'Ball stories' - Attempted help: one boy attempts to break the mirror with the ball, but the ball lands in the bin where it belongs. Accidental help: one boy accidentally breaks the mirror when he throws a ball towards the bin where it belongs.

Each child received one 'push story' and one 'ball story', which varied in terms of whether they describe an accident or a failed attempt. While Cushman et al. (2013) read two stories drawn from the same context, we read two stories drawn from two different contexts, and we did so to avoid confounding problems associated with judging and comparing similar stories.

After each story, children were asked two comprehension probes about the intention and the outcome of the story, and two test questions: "According to you, in the story I just told you, is [character name] a bad boy or not?", and "According to you, in the story I just told you, does [character name] deserve to be punished, or not?" After both stories, we introduced one last procedural change, that is, a choice task with which children were invited to tell which character they think was 'more bad'. We introduced this change to have a further measure of the children's sensitivity to intention.

Results.

Responses to the stories for which a child did not pass both control questions were excluded (age 4 = 10% of responses; age 5 = 3%).

Preliminary analyses on story order effects.

The effect of age on disapproval judgments of accidents or attempts did not interact with story order (all *p*s for interactions > .10). Since children's responses were not affected by story order, we will consider responses to both stories for the following analyses.

Testing the occurrence of an outcome-to-intent shift.

As in the results section of Experiment 1, here we focus our analyses on each child's first response in order to eliminate questions order effects.

The choice measure ('more bad question'). We carried out a logistic regression analysis on children's answers to the question concerning which character was 'more bad', using age group as predictor. We found an age-related increase in pointing to the character who attempted to harm, $\beta = .74$, 95% CI [.53, .95], $z = 7.01$, $p < .001$ (see Figure 1b).

Similarities between choice measure in Experiment 1 and 2. In order to test whether the likelihood to choose who attempted to harm in the choice task was different from the likelihood to choose who attempted to help (Experiment 1), we conducted separate McNemar tests for each age group, and we found no evidence of a difference in the preference for intent-based judgments for help and harm (all *p*s > .38).

Development of badness and punishability judgments. Children's badness and punishability judgments on both types of stories are illustrated in Figure 2b. Conducting a series of logistic regression analyses, using age group as predictor, we found an age-related increase in badness attribution to attempted harm, $\beta = .59$, 95% CI [.33, .84], $z = 4.54$, $p < .001$, and a decrease in badness attribution to accidental harm, $\beta = .89$, 95% CI [.45, 1.33], $z = 3.94$, $p < .001$. There were no age-related changes in punishability attribution neither to attempted nor to accidental harm. Note that 4-year-olds did not attribute badness to attempted harm (39%) more or less often than expected by chance (*p* from binomial test = .256), but their attributions of badness to accidental harm (32%) were below chance level ($p = .047$, binomial test).

Goodness vs. badness and reward vs. punishability. To assess whether children consider negative duties as more restrictive than positive duties, or vice versa, we compared the results from both experiments. Children aged 4 to 6 attributed goodness to attempted help more

often than they attributed badness to attempted harm, age 4: $\chi^2(1, N = 76) = 9.05$; age 5: $\chi^2(1, N = 81) = 8.25$; age 6: $\chi^2(1, N = 80) = 14.05$; all *p*s < .004. This suggests that younger children's responses in the attempted stories are more intent-based when facing helping actions rather than harming actions. Conversely, children's attributions of goodness to accidental help were no more or less frequent than children's attributions of badness to accidental harm. Also the likelihood to attribute deserved reward did not significantly differ from the likelihood to attribute punishability, neither in attempted nor in accidental cases.

Attempted vs. accidental harm. Children aged 5 to 8 judged attempted harm more bad than accidental harm, age 5 and 6: all *p*s from binomial tests < .001; age 7: McNemar $\chi^2(1, N = 39) = 29.26$; age 8: McNemar $\chi^2(1, N = 36) = 30.03$; all *p*s < .001.

Attributions of badness vs. punishment. Cushman et al. (2013) found that the criteria to attribute badness and punishability start to dissociate at 5-6 years. We found that attempted harm was judged more bad than punishable only at 7 years, $\chi^2(1, N = 78) = 13.24$, $p < .001$, and approached significance at 8 years, $\chi^2(1, N = 72) = 5.14$, $p = .023$ (*p* value must be < .01 following the Bonferroni correction). Instead, in each age group, accidental harm was judged less bad than punishable, age 4: $\chi^2(1, N = 76) = 7.62$; age 5: $\chi^2(1, N = 83) = 27.73$; age 6: $\chi^2(1, N = 84) = 14.91$; age 7: $\chi^2(1, N = 78) = 27.19$; age 8: $\chi^2(1, N = 72) = 29.65$; all *p*s < .006.

Testing the constraint hypothesis.

Analysis of mediation effects. Similarly to the case of approvals, we found only a positive correlation between age and badness judgment of attempted harm $r = .39$, $p < .001$, and a negative correlation between age and badness judgment of accidental harm $r = -.27$, $p < .001$. Thus, we failed to replicate Cushman et al. (2013)'s results on mediation effects.

Test question order effects. To facilitate the comparison with Cushman et al. (2013)'s results, the analyses compared three age groups (4, 5, and 6 to 8). Order of questions did not affect younger children's punishability or badness judgments of accidental harm (all *p*s >

.068). By contrast, both older children's (6-8) badness and punishability judgments of accidental harm were affected by questions order (Figure 3b). Badness judgments affected punishability judgments, $\chi^2(1, N = 234) = 4.99$, $p = .026$, and these latter affected badness judgments, $\chi^2(1, N = 234) = 7.95$, $p = .005$. Children who made badness judgment first were less likely to judge the character to be punishable compared with those who made punishability judgment first. The reverse was also true: children who made punishability judgment first were more likely to judge the character to be bad compared with those who made badness judgment first. Thus, each kind of judgment had a constraining effect on the following other judgment, whereas in Cushman et al. (2013), by age 5, badness judgments constrained punishment judgments, but not vice versa.

Discussion.

The results of Experiment 2 confirmed that an outcome-to-intent shift occurs in children's moral evaluation of harming actions. First, when asked to choose the character who is 'more bad', older children pointed to the character who attempted to harm more often than younger ones. Second, with age, the likelihood to attribute badness to the character who attempted to help increased, and the likelihood to attribute badness to the character who harmed accidentally decreased. Third, by age 5, children judged attempted harm as worse than accidental harm.

Cushman et al. (2013) suggested that the shift occurs during preschool years; the criteria to attribute badness and punishability start to dissociate at 5-6 years, when children judge accidental harm less bad than punishable; badness judgment becomes intent-based and punishability judgment remains more outcome-based compared to badness judgment. We found evidence of a shift (see Table 2), and however there are some noteworthy differences in the two pattern of results. In the present study, children judged accidental harm less bad than punishable even at 4 years, but they started to judge attempted harm more bad than

punishable only at 7 years. Also, while the previous study found an outcome-to-intent shift in accidental harm stories only, we found such a shift both in accidental harm and failed attempts stories (see Figure 2b). Moreover, unlike Killen et al. (2011), we did not find any developmental change concerning judgments of punishability. However, we replicated one important aspect of Cushman et al. (2013): older children's punishment attribution was more outcome-based than badness attribution.

Two pieces of evidence supported the hypothesis of a conceptual change in the previous work. First, the intent-based badness judgment mediated the effect of age on punishment judgment. Second, as the shift occurs, children's badness judgment of accidental harm constrained the subsequent judgment of punishment, but not vice versa; children endorsed the criterion 'the character was (not) bad, therefore she deserves (not) to be punished'. We failed to replicate both of these results: 1) we did not find any correlation between age and punishment judgment, and 2) contrary to our expectations, older children's badness and punishability judgments constrained each other. Instead of finding that the new emerging and now central intent-based badness judgment constrains punishability judgment – i.e., a unidirectional constraint – we found that the constraining effect was bidirectional. In sum, while we replicated the previous evidence of an outcome-to-intent shift on badness judgments, and extended such evidence to goodness judgments, we failed to find any support for the predictions derived by the constraint hypothesis.

Procedural and sample differences between studies might be in part responsible for the differences between previous findings and ours. Centrally, our stories were enriched by the description of the harmed (or not harmed) character's emotional reaction. This procedural change was introduced in order to make more salient the harm produced by the main character and it may explain why punishability judgments in the present study and deserved reward in Experiment 1 did not show any developmental change. Here we address the

unexpected result that neither judgments of punishability (Exp. 2) not judgments of deserved reward (Exp. 1) changed with age. Although this is a striking result, we should stress the fact that using a similar experimental procedure in the two experiments lead us to find similar patterns in punishment and deserved reward attribution. However, we need to explain why this pattern of results differs from previous findings (Cushman et al., 2013; Killen et al., 2011).

For instance, studying children between 4 and 8 years, it has been reported an age-related decrease in judging punishable the character who accidentally harmed (Cushman et al., 2013). Also Killen et al. (2011), studying children between 3 and 8 years, reported that younger preschoolers judged acceptable to punish an accidental transgressor, but with age they increasingly judge not acceptable to punish him. In both these studies, children were presented with stories in which the victim's emotional reaction to the accidental harm was not described. Unlike these studies, we told children the emotional state of the character whom is being accidentally harmed (he is 'sad') or helped (he is 'happy').

We argue that the introduction of this slight procedural change might have inhibited older children's intent-based processing and thus increased the likelihood to attribute punishability for accidental harming actions and deserved reward for accidental helping actions. While younger preschoolers' judgments of punishability or deserved reward might be more outcome-based than older children's judgments, the salience of the emotional information might have caused a shift in older children's attention from intentions to consequences, leading them to attribute punishment or deserved reward to an equal rate compared to younger children. Therefore, with respect to previous work, the description of the emotional reaction of the victim or the beneficiary might have biased the responses of older children, that is, the group of participants who would have shown an intent-based judgment of punishability or deserved reward. This shift in attention may be also facilitated

by the common tendency of parents and educational figures to rely on consequences to foster obedience and a quick and stable success in teaching moral rules and social conventional norms (Nucci, 2001; Tindall & Ratliff, 1974).

Two remarks are in order for future research on this issue. First, procedural aspects should be taken very seriously into account when studying children's moral reasoning, since minimal changes threaten the generalizability of the results. Then, further work is needed to clarify under which conditions it is useful to describe the characters' emotional reactions in moral judgment tasks, and to investigate how the evaluation of single actions differs from the evaluation of the agents. Second, future research should be devoted directly to investigate whether the assessment of the emotional state of the character being harmed or helped consistently produced in children a shift from an intent-based to an outcome-based moral judgment. This future work is potentially of great theoretical interest, given the importance that some theories, both in psychology (e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Hoffman, 1991) and in philosophy (e.g., Hume, 1740/1978; Smith, 1759/1948), attribute to the role of empathy in the development and processing of moral judgments.

General Discussion

In Experiment 1, we investigated the outcome-to-intent shift in children's evaluation of helping behavior, analyzing separately attributions of goodness and deserved reward. We reported evidence of a shift occurring in goodness judgment, but we found no evidence for the constraint hypothesis regarding approvals development. In Experiment 2, we replicated previous evidence of an outcome-to-intent shift in badness judgment of harming behavior, but we failed to replicate evidence supporting a conceptual change in the concept of badness (see Table 2 for a summary of these results).

Outcome-to-intent shift in the evaluations of helping and harming agents.

There are three main similarities in the development of our children's judgments of harming and helping actions. First, when asked to choose the character who is 'more good' (Exp. 1) or 'more bad' (Exp. 2), children increasingly showed an intent-based choice, pointing to whom attempted to help or harm, rather than to the agent who accidentally helped or harmed. Second, the likelihood to attribute goodness or badness in failed attempt cases increased with age, and decreased in accidental cases. This pattern of results again highlights an increased sensitivity to intent cues when judging both character's goodness and badness. Third, the outcome-to-intent shift concerned selectively goodness and badness judgments, since neither deserved reward nor punishability judgments showed any developmental change. Although deserved reward and punishability attributions did not show any developmental change, we replicated previous finding by Cushman et al. (2013) consistent with a dual-process theory of moral judgment by reporting that in older children both deserved reward and punishability attributions were more outcome-based than goodness and badness attribution respectively.

There were also some important differences between the development of goodness and badness judgments. In judging the character whose action outcome is accidentally caused, whereas 4-year-olds' attributions of goodness were not yet dissociated from deserved reward attributions, younger children's attributions of badness were already dissociated from punishability attributions (see Figure 2, right panels). Although this difference may support the conclusion that the intent-based badness judgment develops prior to the intent-based goodness judgment when children's evaluation is assessed in accidental cases, focusing on 5- to 8-year-olds' responses to accidents allow us to see that the developments of goodness and badness judgments follow very similar pathways after age 5. In fact, in any age group, no difference was found between the likelihood to attribute goodness to accidental help and the likelihood to attribute badness to accidental harm.

By contrast, a much more marked difference is revealed by a comparison between the development of intent-based goodness and badness judgments, when children are asked to evaluate cases of failed attempts. In judging those cases, 4-year-olds attributed goodness more often than expected by chance, and 5-year-old group was already at ceiling on the goodness attribution measure. As shown in Figure 2 (left panels), children's intent-based badness judgment develops two years later compared to intent-based goodness judgment. Four-year-olds did not attribute badness more often than expected by chance, and only by the age of seven children reached an intent-based badness judgment. Moreover, whereas the intent-based criterion to attribute goodness dissociated from the one the child uses to attribute deserved reward already in the youngest age group, the criteria to attribute badness and punishability started to dissociate only by 7 years of age. In sum, our results suggest that, when we focus on children's evaluations of attempted help and harm, the intent-based badness judgment appears to develop somewhat later than the intent-based goodness judgment.

In the introduction, we asked whether the intent-based goodness judgment would develop prior to, later or simultaneously with the intent-based badness judgment. Our study tentatively suggests that intent-based goodness judgment develops first. When younger children (4-6) heard a story in which someone had a good intention to help or, vice versa, a bad intention to harm, they found relatively easy to attribute goodness to the agents that wanted to help, but for some years they remained unsure in their attribution of badness to somebody who wanted to harm. This suggests that attributing goodness is easier for children than attributing badness. This conclusion is at odds with past research suggesting that the child would learn the notion of what is good somewhat later than the notion of bad (Piaget, 1932; Hill & Hill, 1977; Karniol, 1978; Rhine, Hill, & Wanderuff, 1967; see also Lyn, Franks, & Savage-Rumbaugh, 2008 for a more recent study on language-competent apes).

Our conclusion is also contrary to the prediction that children consider negative duties as more restrictive than positive duties. However, this conclusion may be also seen as consistent with the fact that positive duties are not obligatory, and thus someone who wants to help may be deemed particularly morally good because he could have easily avoided providing help.

One different possibility is that young children show a bias in goodness evaluations due to a default positive assumption: an agent is evaluated positively, as a good agent, unless contradicting evidence is available from his or her actions. This bias will make the generation of responses easier in the case of attempted help as opposed to attempted harm, because only in the latter case the child needs to change her initial response. In evaluating the badness of attempted harm, children have to grow to start taking into account the negative intention, then younger children do not recognize the negative intention as evidence of the character's badness. By contrast, in evaluating cases of attempted help, younger children may be facilitated in showing what appears to be an adult-like intent-based judgment simply by the default assumption that the character is good. An interesting goal for future research will be to investigate the possibility of a 'default positive assumption' in children moral judgments, and to explain why preschoolers find it easier to evaluate helping rather than harming actions.

In sum, for the first time we reported evidence of an outcome-to-intent shift in judging the morality of helping agents. Future research could benefit from analyses that directly compare harming and helping moral evaluations.

Continuity vs. constraint hypothesis.

A secondary aim of our study was to address the problem of continuity in development. Our findings are less consistent with the view that posits conceptual change than with the alternative view that posits conceptual continuity. The former view predicts that by the time the outcome-to-intent shift would occur, the new emerging intent-based judgment should start to constrain the other type of judgment to be also based on intention assessment. The

evidence we collected in Experiment 1 and 2 does not support this view, and rather it is more consistent with the rival view. First, we did not find that the intent-based goodness (or badness) judgment mediates the effect of age on deserved reward (or punishment) judgment. Second, once the shift occurred, children's responses about goodness and deserved reward (or badness and punishability) constrained each other; that is, we did not find a unidirectional order effect. Then, we did not find that goodness or badness intent-based judgment constrained the development of deserved reward or punishment, and we found instead a growing context-sensitivity in judging accidental harm and help cases (Turiel, 1983; Jambon & Smetana, 2013).

On the one hand, preschoolers' goodness judgments of accidental help were constrained by deserved reward judgments. How should we account for these order effects found in young children? Apparently, children endorsed a 'realist' criterion "the agent deserved a reward, therefore she must be good", but their judgments of accidental harm were not affected by questions order. By age 5, however, the developmental pathways in attributing goodness and badness to accidents were similar and similar order effects were found in older children's responses to accidental harm and help.

In sum, the present evidence provides no support for the constraint hypothesis, and suggests that the outcome-to-intent shifts may not reflect changes in the concepts of moral goodness or badness. Such concepts are both outcome- and intent-based from start, and changes in executive function or theory of mind explain the increasing sensitivity to intent cues. In order to provide more direct support for this conclusion, future research should study how the development of intent-based moral judgments is linked to changes in children's performance on theory of mind and executive functions tasks.

Preschoolers vs. infants' early evaluations.

Our conclusion is consistent with recent evidence that comes from infant studies. Using recently developed paradigms to investigate infants' expectations and representations, scholars found evidence not only of an early ability to attribute false beliefs (Baillargeon, Scott, & He, 2010; Low & Perner, 2012), but also of an early intent-based value judgment (e.g., Dunfield & Kuhlmeier, 2010; Hamlin, 2013; Lee, Yun, Kim, & Song, 2015). Some studies used the 'unwilling versus unable' paradigm to demonstrate that infants decide to help based on the assessment of the moral value of the previous intention of those who ask for help. Children are typically presented with actors showing to be either unwilling or unable to please them. In the second year of life, infants prefer to help those who were unable rather than those who were unwilling, but they also show no preference between actors who were able to please them and actors who were unable (Dunfield & Kuhlmeier, 2010). The results pattern from Hamlin (2013) is even more striking. In this latter study, infants' evaluations were tested with the preferential-reaching method. Infants in the first year of life were presented with puppets who attempted but failed or succeeded to either help or hinder a second puppet goal-directed action. Eight-month-olds showed to prefer those who help rather than those who hinder, but relevantly here, they did not showed any preference between those who succeed in their intention and those who simply attempted to produce their intended outcomes. Experiment 3 from Lee et al. (2015) recently expanded on previous research by showing that 12-month-olds already use intention information when inferring others' socio-moral preferences. These results show that even infants' early evaluations are driven by an assessment of agents' intentions.

Future research should be open to a theoretical and empirical work in order to explain the puzzle of why and how a shift from outcome-to-intent occurs in moral reasoning while even infants in their first year of life based their socio-moral evaluations on intention. Positing a conceptual change during preschool years would imply that an already existent intent-based

concept of goodness or badness changes in an outcome-based concept, only to change again during the later preschool years. By contrast, a view that insists on the role of changes occurring outside the moral domain could outline a more parsimonious description of how the moral evaluation develops (Margoni & Surian, 2016a). While an intent-based concept of goodness and badness is already present early in life, and even infants can generate an intent-based implicit evaluation, preschoolers would show a partially outcome-based explicit judgment because of their immature domain-general abilities.

More work could also be done in order to clarify what role and weight culture and social exposure have in the development of an intent-based moral judgment. In the present work, as well as in the vast majority of the studies on the outcome-to-intent shift, only children from large-scale industrialized societies were examined. However, given the emerging evidence of a substantial cross-cultural variation regarding the role of intention in moral judgment in small-scale societies (Clark-Barrett et al., 2016), it would be extremely interesting to test the generalizability of the present results to these societies. In fact, in some small-scale societies (e.g., Hazda from Africa or Yasawa from Pacific Islands), people weigh less the intention and more the outcomes when judging a moral transgression, compared to people from large-scale societies.

This evidence suggests that culture can play a pivotal role in shaping the development of our moral judgment. Both the education explicitly given to children by their parents and the interactions with peers could shape the way children evaluate moral situations. Future research should examine the relative weight of a) culture and social exposure; b) the innate ability to evaluate others' moral behaviors based on an intention assessment (Hamlin, 2013); c) and the development of general abilities that allows the child to generate explicit and controlled evaluations accordingly to his or her implicit conceptual repertoire.

With respect to the latter point, more research is also needed to clarify which particular aspects of executive functioning are central for the development of an intent-based moral judgment. We hypothesize that the development of inhibitory control, occurring during the early years of life, can be deemed responsible for the shift (Margoni & Surian, 2016a). For younger preschoolers that are tested by using elicited-response tasks, to suppress 'wrong' responses based on salient cues such as outcomes, and to select (or set shift to, see Diamond, 2013) 'right' responses based on intention may be too demanding.

Relevance of the present research for clinical psychology and education practices.

Researching on the developmental aspects of intent-based moral judgment promises to lead to important insights for both clinical psychology and pedagogy. First, understanding the developmental shift from outcome-to-intent in moral reasoning can be used to further and better assess the extent of the theory of mind impairment in clinical population such as autistic children. Research on autism has shown that the ability to distinguish between moral and conventional transgressions is spared (Blair, 1996; Zalla, Barlassina, Buon, & Leboyer, 2011), along with a basic moral judgment, that is, the ability to correctly attribute badness or goodness to actions that caused outcomes consistent with agents' intentions (Grant, Boucher, Riggs, & Grayson, 2005; Leslie, Mallon, & DiCorcia, 2006). However, when asked to evaluate cases of accidental harm or attempted but failed harm, adults with high-functioning autism failed to integrate mental state information in their moral reasoning and did not judge accidental harm more permissible than attempted harm (Moran et al., 2011; for a review see Margoni & Surian, 2016b).

Adults with high-functioning autism can pass standard tests for theory of mind (Bowler, 1992; Frith, Morton, & Leslie, 1991) and children with autism can represent intention (Carpenter, Pennington, & Rogers, 2001). However, moral judgment tasks represent a particularly strong test for theory of mind reasoning, above all those that imply the evaluation

of accidental and attempted harming or helping actions. By showing the developmental changes occurring in typical population, scholars will be able to further assess theory of mind reasoning impairments in autistic individuals. Therefore, future research should study also the moral evaluation of cases of accidental or attempted harming/helping actions children with autism.

Second, the current research may have practical implications for educational school programs or parents' caring choices (Nucci, 2001). In fact, knowing at which age children's moral reasoning starts to rely consistently on intention is crucial to improve existing programs of moral or civic education, and to guide parents in a better understanding of when and why their children can fully appreciate the content and the meaning of their moral teaching. In particular, it would be valuable for parents to understand whether their children can understand and benefit from an education that relies on the adults pointing out to the children the mental state quality of the actions instead of the material consequences.

In this respect, it would be of the utmost importance for future research to deepen our understanding of the relationship between the development of the child's moral judgment and the development of his or her moral behavioral tendencies. Little systematic attention has been paid to this complex endeavour (but see Sheskin, Chevallier, Lambert, & Baumard, 2014), that nevertheless has great theoretical and practical importance.

Limitations.

Some limitations of the present study should be pointed out. First, in Experiment 1, children evaluated only helping behaviors. Future research should focus on judgments of different types of positive moral behaviors, such as, for example, the respect for natural entities (i.e., the ecological morality). A related problem was that children evaluated the goodness or the badness (in Exp. 2) of the character by mean of the evaluation of a single action. Many perspectives, such as the social domain theory (Turiel, 1983), center on the

relevance of context for the moral evaluation. Contextual elements, such as the presence of negligence in accidents (see Nobes et al., 2009), may be likely to be inferred when a single action is taken as a proxy for the character evaluation. More relevant for the present study, it can be argued that attributing goodness or badness by evaluating a single action is substantially different than relying on a global evaluation of the individual. In particular, the global evaluation of the character may not relate to the subsequent decision whether the character deserved to be punished or to be rewarded in the same way a narrow evaluation of the single action would do. Therefore, we could predict different questions order effects depending on the evaluation we ask the child to produce.

However, it should also be noticed that stressing the nature of the questions used in our study is not helpful in explaining the discrepancies with the previous work that addressed the issue whether the outcome-to-intent shift reflects a conceptual change or changes occurring outside the moral domain, and that found an age-related change in punishability judgments. In fact, also Cushman et al. (2013) led children to produce a global evaluation of the characters by asking them whether the main character was a naughty/bad boy and not whether the character's action was bad or ok.

Future research should study the evaluation of the character as a product of a broad set of actions (e.g., Feltz & Cokely, 2012), that is, the evaluation of virtue, a supposedly stable quality of the character (Swanton, 2003). For instance, it would be interesting to investigate how children integrate the evaluation of actions belonging to distinct moral categories (such as harm and purity; see Graham, Nosek, Haidt, Iyer, Koleva, & Ditto, 2011) in the global judgment of the character.

Second, although we made an effort to match positive stories from Experiment 1 with negative stories from Experiment 2 along several dimensions, such as the outcomes severity, the wording count, and the general structure of the story (e.g., if the helping story version

tells about a character that ‘stumbles in the carpet and by accident opens the door for his brother’, then the harming version tells about a character that ‘trips on a rock and accidentally pushes his brother over’), future studies may want to use even more similar versions to facilitate the comparison between moral goodness and badness attributions. In particular, one aspect of the story set may arise some skepticism about the validity of the comparison we made between children’s responses to harming and helping cases. It can be pointed out that helping story contexts (‘Tree’ and ‘Door’) depicted both a case of instrumental helping, but harming story contexts were instead more differentiated. A first harming story context depicted a physical harming action toward a person (‘Push’ – pushing the brother to the floor) while a second story context depicted a property destruction (‘Ball’ – breaking the mother’s mirror). However, we argue that despite their differences, the two harming contexts are similar in a fundamental way. In fact, both contexts involve the presence of a victim and some consequences for his or her right and welfare. By describing the emotional reaction of the victim (in the ‘Ball story’ experimenter told that ‘the mother is now sad because her favorite mirror got broken’), we helped the recognition of the victim and the moral transgression that occurred. Therefore, we judged the comparison between children’s responses to harming and helping actions to be appropriate, and not undermined by the particular nature of our story contexts.

A third limitation is that the evidence we reported against a conceptual change is not conclusive. The premise of the main argument for the modular nature of morality was that non-moral elements of children’s developing psychology should not affect selectively the moral concepts of badness/goodness. We found this premise intuitively plausible, but one can challenge it or refute it. Our aim was to assess the extent to which the pattern of results previously reported to constrain theoretical accounts of the locus of the outcome-to-intent shift mirrors a pattern of results concerning children’s approvals. We found a selective

change for goodness and badness judgments, but we did not find the emergence of an intent-based judgment that constrains selectively an outcome-based judgment. Future research should measure children's ToM and executive function skills, and should directly assess the impact of the development of these skills on children's moral reasoning (Buon, Seara-Cardoso, & Viding, 2016; Gvozdic, Moutier, Dupoux, & Buon, 2016; Killen et al., 2011).

A final limitation of this study is that we did not ask children to motivate their judgments and choices. A fascinating goal for future research would be to include also new effective tasks to elicit moral justifications in young children.

Conclusions.

The present study reports, for the first time, an outcome-to-intent shift in the attributions of moral goodness to characters that attempted to help, but failed, or accidentally helped another person. During preschool years, the likelihood to attribute goodness to the character that attempted to help increased, reaching rapidly adults' levels. Also, the absence of good intention is gradually weighted more than the presence of a desirable outcome in evaluating a case of accidental help. The results do not support a conceptual change account of the outcome-to-intent shift: contrary to the conceptual change view, older children's goodness and deserved reward judgments constrained each other. This pattern of results is more consistent with an account that emphasizes conceptual continuity in the development of moral judgment.

CHAPTER 3

Biocentric Moral Reasoning in Preschool Children

This chapter is based on the following original article:

Margoni, F., & Surian, L. (2016). *Biocentric moral reasoning in preschool children.* Manuscript under review.

Abstract

We assessed, for the first time, 4-to-5-year-old children's choices between two contrasting ways of extending ethics to natural entities: anthropocentrism (nature has to be preserved because it helps humans' interests) and biocentrism (nature has to be preserved because of its intrinsic value). Children evaluated the rightness or wrongness of a decision taken by a character acting with either a biocentric or an anthropocentric intention. Children were also asked whether the character deserved a reward or a punishment for having caused, as a side-effect of his actions, an ecological damage or benefit. Preschoolers judged the character who caused accidentally an ecological benefit more worthy of a reward when he had a biocentric intention than when he had an anthropocentric intention, thus showing an emerging preference for biocentrism.

Biocentric Moral Reasoning in Preschool Children

When judging whether an act is morally right or wrong, preschoolers tend to focus on actions' outcome whereas older children increasingly rely on mental state attribution. Since the seminal work of Piaget (1932), researchers showed that children shift from judging moral actions based on outcome to judge them on the basis of intention (e.g., Armsby, 1971; Baird & Astington, 2004; Costanzo, Coie, Grumet, & Farnill, 1973; Killen, Mulvey, Richardson, Jampol & Woodward, 2011; Moran & O'Brien, 1983; Nobes, Panagiotaki, & Pawson, 2009; Yuill, 1984). According to recent evidence, children's verbal moral judgment starts to be based mainly on intention rather than action outcome around the age of five (Cushman, Sheketoff, Wharton & Carey, 2013; Margoni & Surian, 2017). However, whether and how intentions that are associated with different moral views affect preschool children's moral judgment remains unclear.

It is now widely acknowledged that a fundamental part of morality concerns ecological issues. Should we integrate also the well-being of natural entities (such as non-human animals or other living beings) in our moral scope? If yes, in which way should we pursue this goal and on what grounds? Nowadays, these are pressing questions. Moral psychologists have recently begun to study people's moral reasoning about ecological issues in adults (e.g., Clayton, 1998; Corraliza, Collado, & Berthelmy, 2013; Gagnon Thompson & Barton, 1994; Kaiser, Ranney, Hartig, & Bowler, 1999; Kortenkamp & Moore, 2001) as well as in children (e.g., Howe, Kahn, & Friedman, 1996; Hussar & Horvath, 2011; Kahn, 1997; Kahn & Friedman, 1995; Kahn & Lourenco, 2002; Kellert, 1985). Overall, the studies revealed that even first-graders value the relationship with the natural environment and consider environmental harm a violation of a moral obligation.

In the environmental psychological literature, biocentrism is distinguished from anthropocentrism as a different way of reasoning about the extension of ethics to nature

(Kahn & Friedman, 1995). According to the anthropocentric view, nature is valued because how it is treated affects humans' interests. By contrast, for biocentrism nature has to be valued because of its intrinsic value. By relying on the distinction between these two ethical orientations, we differentiate between two intentions with which an agent can act toward a natural entity: an anthropocentric and a biocentric intention. A well-known example of the anthropocentric view can be found in the Genesis book, where we read that God created the natural world with all its animal inhabitants for men's benefit and rule (see White, 1967). In the Book of Job, where the human being has a decentered position within creation, or in St. Francis' Canticle of the Creatures, one may instead find early examples of the biocentric view.

The Development of Biocentric Reasoning in School-ages Children

A number of studies investigated anthropocentric and biocentric reasoning, both in adults (e.g., Casey & Scott, 2006; Gagnon Thompson & Barton, 1994; Kortenkamp & Moore, 2001; Milfont & Duckitt, 2010; Schultz, 2000; Snelgar, 2006; Stern & Dietz, 1994) and in children (e.g., Hussar & Horvath, 2011; Kahn & Friedman, 1995; Kahn & Lourenco, 2002; Kortenkamp & Moore, 2009). Overall, the studies revealed that 6-year-olds already assume a moral obligation towards natural entities, but the biocentric reasoning does not develop until late childhood (Kahn, 1997; Kahn & Friedman, 1995; but see also Hussar & Horvath, 2011). School-aged children use the anthropocentric reasoning more often than the biocentric reasoning, which is indeed more commonly found in older children, not before the fifth grade.

Kortenkamp and Moore (2009) investigated age-related changes in the use of intention information (anthropocentric vs. biocentric helping intention) in shaping school-aged children and adolescents' moral judgments of ecological damage. Children and adolescents judged an agent that caused an ecological damage to be less blameworthy if the intention was described

as biocentric rather than anthropocentric. That is, having a biocentric intention lessen the condemnation. The study also reported some evidence that younger participants (fifth-graders) do not judge differently agents with biocentric intention and agents with anthropocentric intention when interviewed about a moral scenario that did not involve concerns about animals, but this was not true when children were interviewed about actions concerning animal welfare.

Other studies investigated the understanding and use of biocentric reasoning from 6 to 10 years mainly reporting than children's reasoning focuses on anthropocentric concerns (Kahn & Friedman, 1995). However, no research to our knowledge investigated whether preschoolers understand and use biocentric reasoning. By using a new version of the Kortenkamp & Moore (2009)'s task, we aimed to assess whether children younger than 6 years show a preference for the biocentric over the anthropocentric reasoning.

The Current Study

For the first time, we investigated whether preschoolers hold a preference for biocentric over anthropocentric intention information in judging the rightness or wrongness and the deserved reward or punishment of both ecological damage and ecological benefit. Our study introduced some novelties: it is the first (a) to examine children younger than 6 years and (b) to focus on children's evaluation of both ecological damage and ecological improvements.

It is worth asking whether some developmental changes in preferring biocentric to anthropocentric intentions occur prior to the primary school period. Studies on the outcome-to-intent shift in moral judgment found that children at the age of five start to rely on mental state information when judging moral cases (e.g., Cushman et al., 2013; Margoni & Surian, 2017). Given that 5-year-olds judge moral cases based on intentions, we asked whether their judgments rely also on different types of intentions.

We presented children with moral scenarios in which a character took an altruistic decision with either an anthropocentric or a biocentric intention that, however, as a side-effect, caused an ecological damage or benefit. If children prefer biocentric to anthropocentric intentions, we should find that their judgments of ecological damage are harsher and their judgments of ecological benefit are more favorable when the character's intention is biocentric rather than anthropocentric.

Although the vast majority of studies on moral judgment focused on how children and adults evaluate transgressions, here we were also interested in how children judge actions that brought about a desirable outcome. So, for the first time, we investigated whether there is an effect of intention type (biocentric vs. anthropocentric) on preschoolers' evaluation of decisions resulting in a benefit for the natural environment.

Following Kortenkamp and Moore (2009; see also Coleman & Temple, 1996), children were presented with scenarios involving damages or benefits to non-human animals (birds). Past studies provided some evidence that children understand better the moral difference between biocentric and anthropocentric intentions when they evaluate scenarios involving non-human animals rather than plants or other aspects of nature (e.g., a shoreline or a park) (Kahn & Lourenco, 2002; Kortenkamp & Moore, 2009). Therefore, we presented children with a scenario used by Kortenkamp and Moore (2009), which originally involved a cat owner who decided to let his cats out of his farmhouse with an anthropocentric intention (the cats will wreck house furniture) or a biocentric intention (the cats need to have fun outside) and, as a results, the cats killed some birds.

Method

Participants. Participants were 38 preschoolers ranging in age from four to five years (M = 65 months, SD = 5 months, 19 girls). Another three children were excluded because they were distracted during the experiment. Children were recruited in three nursery schools

nearby Trento, in Italy. All children were Italian native speakers and no one presented sensory or cognitive impairments. All the parents gave their informed consent on behalf of their children, and the University Ethics Committee approved the experimental procedure.

Materials and procedure. All children were presented individually with four illustrated stories in a within-subjects experimental design. Each story depicted a character acting with a specific intention, either biocentric or anthropocentric, and his action resulted in an outcome that either helped or harmed a natural entity. The stories were drawn from a single story context, while intention (biocentric vs. anthropocentric) and outcome (help vs. harm) were orthogonally manipulated. The scenario we used is reported below (the manipulation of intentions and outcomes are in brackets):

Fred spends a holiday in a friend's farmhouse. Fred has two very active cats. During the day, Fred lets the cats out of the house so they [can have fun in the meadows and he thinks it makes them happy to be outside / do not wreck his friend's house furniture and he is also happy if the house remains tidy]. Around the farm, some birds nested in the trees. Since Fred moved to the farm, his cats [have saved the lives of some little birds that lived there by scaring a big naughty bird that wanted to kill the little birds / have killed some little birds that lived there].

Stories were read out loud to the children and the reading was accompanied with illustrations (see Fig. 1 for an example). Illustrations were used to alleviate memory constraints and were placed in front of the child one by one at the appropriate time during the reading. The last illustration of each story remained in front of the child for reference during the moral evaluation task.

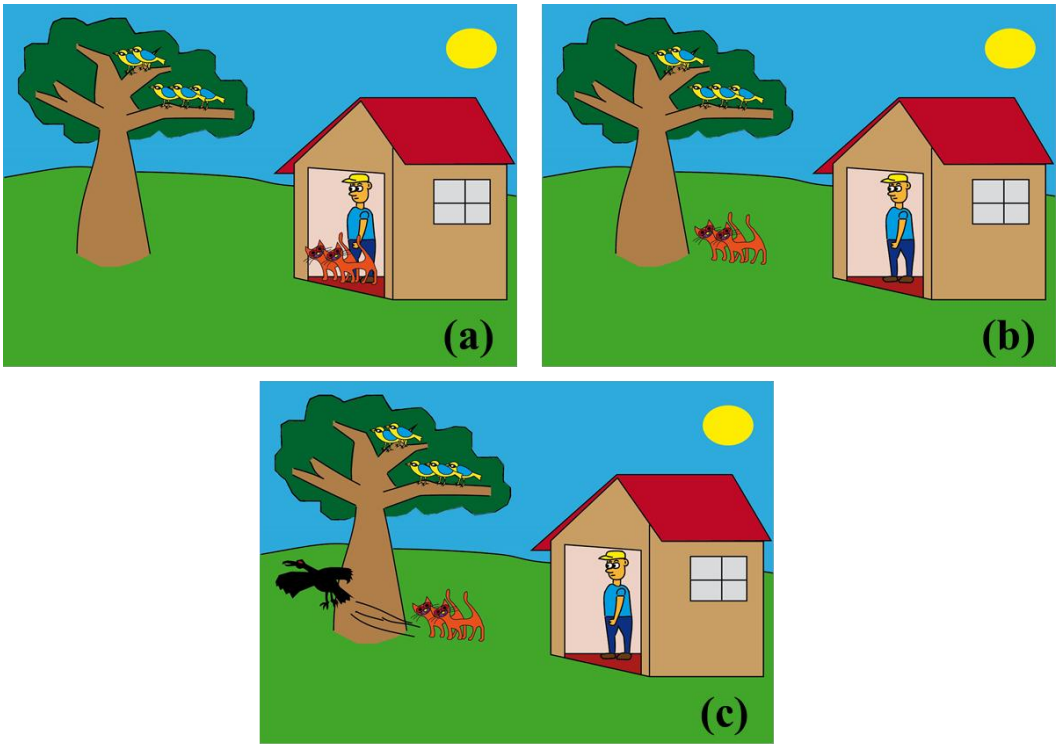

Figure 1. Illustrations depicting a character with a biocentric or anthropocentric intention (deducible only from the story text) which decision determined a positive side-effect. Each illustration was accompanied with a text: (a) Fred spends a holiday in a friend's farmhouse, and he has two very active cats. (b) During the day, Fred lets the cats out of the house so they [can have fun in the meadows and he thinks it makes them happy to be outside/ do not wreck his friend's house furniture and he is also happy if the house remains tidy]. (c) Around the farm, some birds nested in the trees. Since Fred moved to the farm, his cats have saved the lives of some little birds that lived there by scaring a big naughty bird that wanted to kill them.

After each story, children were asked two yes-no control questions, on whether the character wanted to let the cats out of the house and on whether the cats caused the outcome described in the story. With these questions, we assessed children's comprehension of the story content and we focus children's attention on the relevant information (intention and

outcome). Children were then asked two questions: (a) “According to you, was it right or wrong that Fred let the cats out?” – “How much was it [right/wrong]? A little or very much?”; (b) “According to you, does Fred deserve a reward or a punishment for having let the cats out?” – “How much should we [reward/punish] him? A little or very much?” To help children answering, we presented them with a response scale containing images (smiley o sad faces) as anchors (Reynolds-Keefer, Johnson, Dickenson, & McFadden, 2009). Therefore, each response can be processed on a four-point scale anchored at 0 with *very wrong/punishable*, at 1 with *a little wrong/punishable*, at 2 with *a little right/rewardable*, and at 3 with *very right/rewardable*. Stories presentation was randomized using one of the resulting four Latin-square orders, and the order of control and test questions was counterbalanced.

The story context we used was modeled on the ‘free-ranging domestic cat owner’s dilemma’, which has been used in previous works (Coleman & Temple, 1996; Kortenkamp & Moore, 2009). However, we modified it in two relevant ways: (a) in our study, some versions included a positive side-effect (birds been helped) rather than a negative side-effect (birds been harmed); (b) the cat owner was not the house owner, so that his anthropocentric intention to avoid wrecking the furniture cannot be clearly ascribed to egoistic motives. Since the original scenario was affected by the problem that the anthropocentric intention could also be construed as an egoistic intention (the cat owner was also the house owner), we modified the story and children were told that the cat owner was a guest at a friend’s house.

Results

The responses to the stories where the child failed at least one comprehension probe (2.6% of the responses) or the experimenter made an error in assessing the child’s judgments (3.9%) were excluded from analyses. Preliminary analyses revealed no significant effects of

children's sex or questions' order on moral judgments, *F*s > 1. The data were therefore collapsed across these factors in the subsequent analyses.

An ANOVA with intention (biocentric, anthropocentric) and outcome (positive, negative) as within-subjects factors revealed only a significant main effect of outcome (birds saved vs. birds killed) on both rightness or wrongness, $F(1, 31) = 128.29$, $p < .001$, $\eta^2 = .81$, and deserved reward or punishment judgments, $F(1, 31) = 106.95$, $p < .001$, $\eta^2 = .75$. As expected, when the outcome was positive children judged the character's decision more right and the character more worthy of a reward than when the outcome was negative.

Judgments of rightness and deserved reward of positive scenarios were more extreme ($M = 2.7$, $SD = .63$; i.e., near Likert point scale 3 = *very much right/rewardable*) compared to judgments of wrongness and punishability of negative scenarios ($M = .77$, $SD = .96$; i.e., near Likert point scale 1 = *a little wrong/punishable*). It may be suggested that positive scenarios were better understood relative to their valence component than negative scenarios. In fact, 6.7% positive scenarios evaluations were negative (Likert point scale 0 or 1), but 18.6% negative scenarios evaluations were positive (Likert point scale 2 or 3).

Further planned analyses revealed an intention (biocentric, anthropocentric) effect on deserved reward judgments, $F(1, 36) = 4.93$, $p = .033$, $\eta^2 = .12$ (Table 1 and Fig. 2). When the side-effect of the action was positive, children judged the character with the biocentric intention to be more worthy of a reward than the character with the anthropocentric intention. However, children did not judge the decision taken with biocentric intention more or less right than the decision taken with anthropocentric intention. Also, we did not find any effect of intention on wrongness and punishability judgments, all *p*s > .74.

	INTENTION							
Scenarios:					Biocentric		Anthropocentric	
POSITIVE	df	*F*	*P*	η^2	M	SD	M	SD
Rightness	1, 36	0	1	0	2.70	.66	2.71	.65
Reward	1, 36	4.93	.033	.12	2.78	.53	2.61	.68
NEGATIVE								
Wrongness	1, 32	.11	.74	.004	.67	.99	.74	.90
Punishment	1, 32	.03	.86	.001	.82	.95	.85	1.02

Table 1. Effects of intention on rightness or wrongness judgments and deserved reward or punishment judgments.

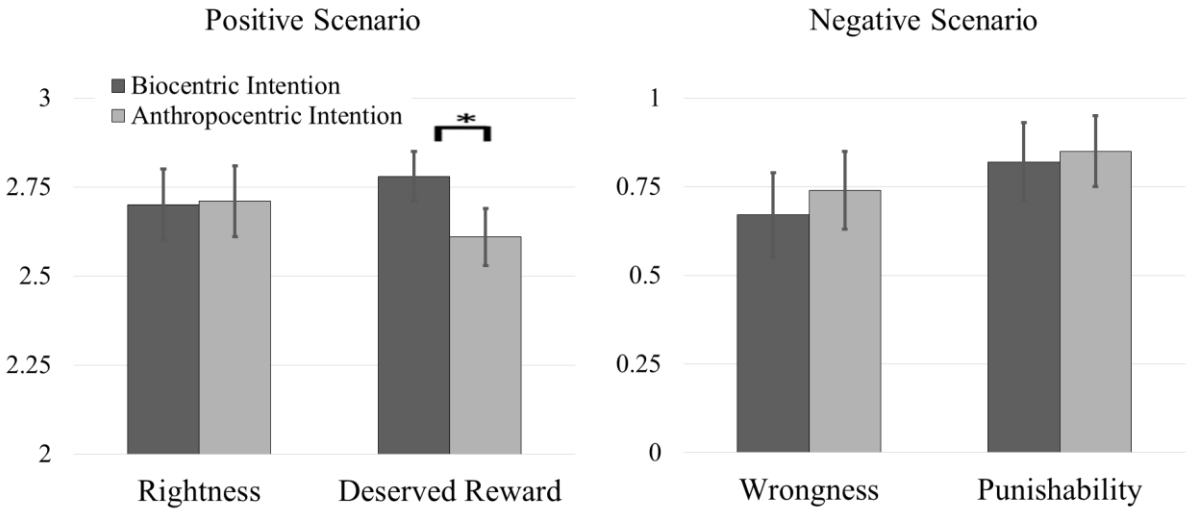

Figure 2. Children's moral judgments of decision rightness or wrongness and decision maker's deserved reward or punishment for positive (a) and negative scenarios (b), on a scale anchored at 0 with *very wrong/punishable,* at 1 with *a little wrong/punishable,* at 2 with *a little right/rewardable,* and at 3 with *very right/rewardable*. Error bars show the magnitude of the standard error. * $p < .05$.

Discussion

For the first time, we investigated whether preschoolers evaluate differently biocentric and anthropocentric motives when they have to judge actions that brought about, as a side-effect, either an ecological damage or an environmental improvement. We asked whether children's judgments about the rightness or wrongness of an agent's action and his deserved reward, or punishment, are differentially affected by the agent's biocentric and anthropocentric intentions. This evidence may help to address the issue of whether preschoolers endorse a biocentric or an anthropocentric reasoning about the inclusion of natural entities within the 'moral circle' (Singer, 2011).

Overall, the results of the current study suggest that preschoolers show a weak but significant preference for biocentric intentions. In fact, children's judgments of deserved reward of the character that caused a side-effect ecological benefit were affected by the intention type: they judged the decision maker more worthy of a reward when he acted with a biocentric intention rather than an anthropocentric intention.

Although a number of studies already suggested that at 5 years preschoolers' verbal moral judgments start to rely on mental state information (e.g., Cushman et al., 2013; Margoni & Surian, 2017), here we reported that they show an emerging preference between two types of intentions (biocentric vs. anthropocentric) that are relevant to adults' reasoning about ecological issues. This preference emerged in the evaluation of actions that produced an ecological improvement, but was not found in the judgments of actions that generated an ecological damage. This result is consistent with past studies showing that biocentric moral reasoning does not fully develop until later childhood, and a full understanding of the moral difference between biocentric and anthropocentric intentions is achieved only during or even after the primary school years (Kahn & Friedman, 1995; Kellert, 1985; Kortenkamp & Moore, 2009).

The finding that the manipulation of intention affects selectively deserved reward judgments is consistent with a recent result in the study of children's intention-based moral judgments of helping actions (Margoni & Surian, 2017). It may be suggested that at this age, when evaluating (accidental) helping outcomes, children's attribution of deserved reward is a more reliable index of their moral approvals than judgment of goodness or rightness (Piaget, 1932; Kohlberg, 1969). This, in turn, can help explaining why here we found a preference for biocentric motives in children's attributions of deserved reward but not in children's judgments of action rightness.

It is not clear why we detected an effect of intention manipulation in the positive scenarios evaluations but not in the negative scenarios evaluations. One possibility is that while in the positive scenario the outcome (help) was matched for valence with the decision maker's intention (to help or to respect), in the negative scenario the outcome (harm) was not matched for valence with the (neutral) intention. Since the intention was neutral but the outcome was negative, children's computational processing in the negative case may have been more demanding compared to the processing underpinning children's evaluations of positive scenarios. Thus, in turn, processing demands may have hindered children's ability to show verbally their preference for biocentrism in their judgments of punishability.

Limitations and conclusion.

A limitation of the current study is that we presented children with only one scenario context that is a modified version of the 'cat scenario' (Coleman & Temple, 1996; Kortenkamp & Moore, 2009). Although we ameliorated the scenario in order to avoid the overlapping between egoistic and anthropocentric motives, and we chose a scenario involving animals, which is the best salient ecological situation for children at this age, besides having assessed children's judgments of ecological benefits along with damages, a problem of generalizability to different scenario contexts still affects our results. Future studies should

investigate children's judgments with tasks employing different and new moral scenario contexts. Currently, a priority for the research on ecological moral judgment is the proper design of the stimuli, which is a methodological but, at the same time, a conceptual challenge.

Another promising direction for future research might be to investigate whether children raised in different, rural or non-Western cultures endorse a more biocentric moral reasoning than Western children do. We showed that preschoolers do not fully weight the difference between biocentrism and anthropocentrism in their moral judgments, although past research have showed that older children do it (Kortenkamp & Moore, 2009). By contrast, cultures or groups that offer to their members a close interaction between man and natural environment may foster very early in life the development of a biocentric moral reasoning.

In conclusion, for the first time, the present study investigated preschoolers' moral preference between two main types of ecological reasoning, namely anthropocentrism and biocentrism. We studied 4- and 5-year-olds' judgments, and we reported evidence that children at this age show an emerging preference for biocentric intentions.

Nowadays, ecological problems are increasingly central for the survival of our and other species. It should then be regarded as useful and right to educate the new generations toward the development of a morality the scope of which includes also the well-being of non-human living entities. Children are likely to have a spontaneous tendency to reason accordingly with a biocentric view, and very early they may be willing to reason about nature according to the principle "Even if there's no rules you should respect ... (and) be good to the environment." (Hussar & Horvath, 2011). However, education may play a crucial role in consolidating, refining or inhibiting this tendency. A fundamental step in the development of educational interventions will be to clarify what younger children can understand, what heuristics biases

inform their moral reasoning, and which mechanisms underpin the development of their reasoning about natural entities.

CHAPTER 4

Explaining the U-shaped Development of Intent-based Moral Judgments

This chapter is based on the following original article:

Margoni, F., & Surian, L. (2016). Explaining the U-shaped development of intent-based moral judgments. *Frontiers in Psychology,* 7:219. Doi: 10.3389/fpsyg.2016.00219

Abstract

When preschoolers evaluate actions and agents, they typically neglect agents' intentions and focus on action outcomes instead. By contrast, intentions count much more than outcomes for older children and adults. This phenomenon has traditionally been seen as evidence of a developmental change in children's concept of what is morally good and bad. However, a growing number of studies shows that infants are able to reason about agents' intentions and take them into account in their spontaneous socio-moral evaluations. Here we argue that this puzzling U-shaped trajectory in children's judgments is best accounted for by a model that posits developmental continuity in moral competence and emphasizes the effect of immature executive function skills on preschoolers' performance.

Explaining the U-shaped Development of Intent-based Moral Judgments

Mental state reasoning is required in several tasks, from inferential communication and the interpretation of social situations to the socio-moral evaluation of actions and agents. Children at an early age start to accuse peers by crying loudly "you did it on purpose!", and legal systems typically distinguish harmful acts that are performed intentionally from acts that accidentally produce personal harm. A large body of developmental research investigated when and how children acquire the ability to attend to agents' intentions and action outcomes in their socio-moral judgments, but the conclusions one can draw from infant studies seem at odds with the conclusions one can draw from studies on older children. Infants seem to possess abilities that young preschoolers' responses do not reveal. In the present work, we address this puzzle by first reviewing relevant results on socio-moral reasoning in infants and children. Then, we evaluate different proposals put forward to explain the reported developmental changes and the apparent contradiction between infants and preschool children's responses.

The Outcome-to-Intent Shift in Preschoolers' Moral Reasoning

Since Piaget's (1932) seminal work, a large body of studies has shown that a crucial developmental change from an outcome- to an intent-based moral evaluation occurs in the late preschool years. A typical Piagetian task would consist of evaluating which of two characters is more naughty and deserves to be punished. Piaget presented children with a story in which a supposedly well-intentioned character accidentally caused serious material damage (e.g., he broke 15 cups), and another story in which a bad-intentioned character caused, also by accident, less serious damage (e.g., he broke one cup). Younger children (aged 6-7) judged the character that produced serious material damage to be more naughty and punishable, whereas older children judged the bad-intentioned one to be more naughty and punishable. These and other similar findings were taken as evidence of a shift from an

initial outcome-based ('objective') moral judgment to a later intent-based ('subjective') moral judgment.

Subsequent research overcame several methodological limitations of Piaget's work (Farnill, 1974; Karniol, 1978; King, 1971; Nelson, 1980), but confirmed the occurrence of an outcome-to-intent shift. During the 1970's and 1980's, many different tasks were developed to investigate this phenomenon. By reducing the cognitive processing necessary to answer experimenters' questions, scholars found that even preschoolers, at age 3, can attend to agents' intentions in their moral evaluations (e.g., Armsby, 1971; Farnill, 1974; Yuill, 1984; Yuill & Perner, 1988). Nevertheless, Piaget's main claim concerning the outcome-to-intent shift found further support, since children older than 4-5 years relied more on intention and less on outcome, whereas younger children showed the opposite pattern (e.g., Baird & Astington, 2004; Costanzo, Coie, Grumet, & Farnill, 1973; Cushman, Sheketoff, Warton, & Carey, 2013; Imamoglu, 1975; Keasey, 1978; Moran & O'Brien, 1983; Nobes, Panagiotaki, & Pawson, 2009; Wainryb, Brehl, & Matwin, 2005).

When intentions and outcomes lead to conflicting responses, as in the cases of *failed attempts to harm* and *accidental harm*, young preschoolers attend to outcome more than older children, relying mostly on outcome (e.g., Helwig, Hildebrandt, & Turiel, 1995), or equally on intention and outcomes (e.g., Cushman et al., 2013; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011). With age, the condemnation of attempted but failed harm increases (Helwig et al., 1995), whereas the condemnation of accidental harm decreases (Cushman et al., 2013; see also Killen et al., 2011 on the development of an intent-based punishability evaluation).

While intentions dominate adults' attribution of moral goodness and badness, adults often rely on both intent and outcomes to evaluate the punishability of agents (Cushman, 2008). A recent dual-process model explains why this is so: adults' moral reasoning is

generated by the work of two independent and sometimes conflicting processes, one that attributes value to actions and assesses agents' mental states, and the other that evaluates the causal responsibility for action outcomes (Cushman, 2013; Cushman, 2015). This proposal contradicts the Piagetian view, which posited a full replacement of the outcome-based judgment by an intent-based judgment.

Infants' Intent-based Socio-Moral Evaluations

Extrapolating the developmental trajectory found in preschoolers, one may predict that infants and toddlers would rely mostly on action outcome rather than agents' intention, assuming that they can produce a moral judgment. However, recent evidence shows that this is not the case. Several studies suggest that, in the first year, infants are able to distinguish between intentions and outcomes, they evaluate helping, harming and distributive actions, and they rely, for these evaluations, on intentions rather than outcomes.

Experimental studies used both elicited-response tasks and spontaneous-response tasks in the *violation-of-expectation paradigm*. Both research strategies found that, by the end of the first year, infants are able to attend to agents' intentions and understand successful as well as failed actions (e.g., Brandone & Wellman, 2009; Gergely & Csibra, 2003; Woodward, 1998). This early understanding of failed attempts generalizes to first- as well as third-party socio-moral evaluations (Behne, Carpenter, Call, & Tomasello, 2005; Dunfield & Kuhlmeier, 2010; Hamlin, 2013; Lee, Yun, Kim, & Song, 2015).

In studies on *first-party* evaluations, infants were engaged in interactions with an experimenter and were presented with actors that were either unwilling or unable to please them (e.g., Behne et al., 2005; Dunfield & Kuhlmeier, 2010; Marsh, Stavropoulos, Nienhuis, & Legerstee, 2010). While the outcomes were identical in both conditions, intentions were different (negative for 'unwilling agents', positive for 'unable agents'). Infants responded differently to these two cases, showing that they used intention cues to guide their *first-party*

evaluations and preferences. Nine-month-olds' spontaneous signals of impatience (such as reaching and banging or looking and turning away) revealed that they become more agitated when they interact with actors unwilling to provide them with a toy (Behne et al., 2005). Moreover, using a manual choice measure (infants have to choose between two contrasted individuals), some studies found that by the second year of life, infants choose to help an unable over an unwilling actor, when asked to help someone. By contrast, infants were equally likely to help able agents, who displayed positive intentions and successful actions, and unable agents, who displayed positive intentions and unsuccessful actions (Dunfield & Kuhlmeier, 2010). Overall, these studies show that infants process information about intention and use it to evaluate others' behaviour.

Further studies on infants' representations of harm and help examined *third-party* socio-moral evaluations (e.g., Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Hamlin & Wynn, 2011; Hamlin, Wynn, & Bloom, 2007; Kuhlmeier, Wynn, & Bloom, 2003; Meristo & Surian, 2014). Infants observed events in which an agent either helps or hinders the goal-directed action of another agent. Their evaluations of prosocial and antisocial actors were typically tested using a manual preference task. Early in their first year of life, infants consistently prefer the helper over the hinderer (Hamlin et al., 2007; Wynn, 2008). Also, at 16 months they prefer agents performing fair over unfair distributive actions (e.g., Geraci & Surian, 2011). When evaluating an agent's behavior, infants are able to take into account not only a person's intention, but also other relevant mental states such as informational states and beliefs (Choi & Luo, 2015; Hamlin et al., 2013; Meristo & Surian, 2013; for a review: Baillargeon, Scott, He, Sloane, Setoh, Jin, Wu, & Bian, 2015).

Hamlin (2013; see also Hamlin et al., 2013) played a puppet show in which puppets either try but fail or succeed to help (or hinder) someone's goal-directed action. Eight-month-olds preferred a helper (failed or successful) over a hinderer, but, most importantly here,

infants did not prefer the successful helper (displaying both intention and relevant outcome) over the puppet that attempted to help, but failed (showing a good intention, but no relevant outcome). This suggests that infants' preferences were guided by agents' intentions rather than outcomes. Moreover, studying expectations by measuring spontaneous looking behavior, scholars recently found that by the end of the first year, infants infer agents' socio-moral preferences by taking into account the agents' information about others' prosocial and antisocial intentions (Lee et al., 2015). They expect that an agent would prefer to approach a second agent who has previously shown a good intention, no matter what the consequences of the second agent's action were.

How can we reconcile the classic results of preschoolers' outcome-to-intent shift with these recent results of infants' intent-based expectations and evaluations? A lesson may be learned from the literature on theory of mind.

How to Account for Seemingly Conflicting Results: The Case of False Beliefs Tasks

The description of the intent-based judgment development sketched above, that is, an initial intent-based evaluation developing from an outcome-based evaluation that in turn shifts again towards an intent-based evaluation, resembles the '*puzzle about belief*' (Perner & Roessler, 2012), that is, the puzzle regarding the development of Theory of Mind. Using traditional elicited-response tasks to study children's attributions of false beliefs, researchers initially concluded that the ability to attribute false beliefs does not emerge until about the fourth birthday (e.g., Baron-Cohen, Leslie, & Frith, 1985; Wellman, Cross, & Watson, 2001; Wimmer & Perner, 1983). However, using *violation-of-expectation* and *anticipatory-looking* spontaneous-response tasks, researchers began to study also infants' mentalizing abilities. Using these and others tasks, scholars demonstrated that babies at least in their second year of life are able to attribute reality congruent and incongruent mental representations across several situations (Baillargeon, Scott, & He, 2010; Buttelmann, Carpenter, & Tomasello,

2009; Low & Perner, 2012; Luo, 2011; Southgate, Senju, & Csibra, 2007; Surian, Caldi, & Sperber, 2007; Surian & Geraci, 2012).

Why do 3-year-olds fail to attribute false beliefs when their abilities are tested on elicited-response tasks? There are two possible answers to this question. First, one may posit continuity during development and argue that preschoolers fail because they do not have the necessary executive function skills to pass an elicited-response task. Second, one may posit a conceptual change during development and argue that the representations and processes involved in resolving spontaneous-response tasks are fundamentally different from the ones involved in resolving elicited-response tasks.

Young preschoolers may succeed in representing the agent's false belief, as infants do, but fail to select the right response and inhibit the wrong response when they are questioned via an elicited-response task (Baillargeon et al., 2010; Leslie & Polizzi, 1998). An *innate modular account* posits that from an early age babies are able to represent and use others' mental states to understand social situations (Kovács, Téglás, & Endress, 2010; Leslie, German, & Polizzi, 2005; Surian et al., 2007). What really develops is the set of cognitive abilities that children need to exploit their representational skills. At 3 years, executive function skills are not sufficiently developed to meet the processing demands of the elicited-response tasks (Thoermer, Sodian, Vuori, Perst, & Kristen, 2012). The continuity account is receiving growing experimental support, but it is still controversial. Many argue for a *conceptual shift account* according to which infants that pass a spontaneous-response task show a qualitatively different level of understanding compared to children who pass an elicited-response task (e.g., Wellman, 2014).

Reconciling Results on Infants and Children at the Processing Level

As in the literature on false-belief understanding, we can draw a distinction between two main positions. First, an *emergence* view posits that during preschool years a conceptual

change occurs in moral competence. The construction of a novel conceptual competence explains why school-aged children's judgments differ from preschoolers. This view does not deny the role of executive function skills, as these are certainly involved in theory construction and revision processes (Cushman et al., 2013). Second, an *expression* view posits conceptual continuity during development and sees the role of executive function in a very different way. It claims that developmental differences result solely from changes in executive function, or theory of mind, that are external to the moral competence (Chandler, Sokol, & Hallett, 2001; Killen et al., 2011; Zelazo, Helwig, & Lau, 1996). We argue that the studies on infants' spontaneous socio-moral evaluations we briefly reviewed above favor the latter view and challenge the former, assuming that a "rich interpretation" (Aslin, 2000) of the infant studies is the correct one. Studies on infants' evaluations suggest that infants can employ an intent-based concept of moral goodness and badness in their socio-moral evaluations. Therefore, the development of intent-based moral judgment is unlikely to derive from a conceptual change occurring in the preschool years.

If infants already possess an intent-based concept of moral badness and goodness, can executive limitations account for preschoolers' outcome-based judgments? The expression view claims that young preschoolers fail at weighting intentions more than outcomes because of processing demands of the task. The additional processing demands of the elicited-response tasks compared to the spontaneous-response task used in the infant literature, lead kindergartners to produce outcome-based evaluations. With the acquisition of sufficient executive function (roughly at 4), children's responses on elicited-response tasks can gradually match infants' spontaneous ones, and become mostly intent-based. When judging an action or an agent in elicited-response tasks, for preschoolers it is difficult to suppress cues concerning action outcomes, while older children may have the sufficient executive function abilities to inhibit an outcome-based judgment and select an intent-based response (or *set*

shift to an intent-based response, see Diamond, 2013; Miyake, Friedman, Emerson, Witzki, Howerter, & Wager, 2000).

Highlighting how different tasks tap different forms of evaluation, and distinguishing between elicited and spontaneous responses may provide the key to solving the *puzzle about intent-based moral judgment*, and avoiding two conclusions that appear highly implausible (Hamlin, 2013). First, it would be implausible that an early tendency to privilege intentions over outcomes emerges during infancy only to be replaced during preschool years by the opposite tendency to privilege outcomes over intentions, or to weight these cues equally, and eventually be again replaced with a final tendency to privilege intentions. Second, it would be also very odd to posit that infants' evaluation system is not related to the later evaluation system, so that we would have two intent-based moral evaluations mutually independent. Conversely, it is likely that if an evaluation system emerges, it will not be replaced later in the development by a new system that serves the exact same function. In order to test directly the expression account, future research should also investigate young preschoolers' generation of intent-based moral evaluations in spontaneous-response tasks.

Additional Factors That May Affect the Outcome-to-Intent Shift

In an expression view, several internal and external factors may promote the emergence of an intent-based elicited response. Among the internal factors, we can include the frontal lobe maturation underlying the acquisition of executive functions (Benes, 2001; Huttenlocher & Dabholkar, 1997; Miller & Cohen, 2001; Moriguchi, 2014; Moriguchi & Hiraki, 2009; Moriguchi & Hiraki, 2011; but see also Knight & Stuss, 2002; Lepsien & Nobre, 2006). Among the plausible external factors, one could include interactions with adults and peers (e.g, Tomasello, Carpenter, Call, Behne, & Moll, 2005). Preschoolers start to be considered somewhat responsible for their actions by their parents, and parents correct their behaviors by pointing to actions outcomes (Piaget, 1932) or negligence (Nobes et al., 2009). However, this

may not be true for infants and older school-aged children. While infants' actions outcomes are limited in their valence and severity, and parents do not deem their children fully responsible for what they cause, older children develop a more controlled behavior, and parents or peers now privilege a comprehensive evaluation of children's intentions and outcomes.

Now, moving from an explanation concerning proximal causes (the processing level discussed in the previous section), to an explanation concerning distal causes (the evolutionary level), one promising perspective is offered by the *life-history theory*. Life-history theory is an approach in evolutionary biology that seeks to explain the timing of the organism's ontogenesis by linking it to relevant evolutionary pressures (Kaplan & Gangestad, 2005). The emergence of a trait in the phenotype has both costs and benefits for the organism with regards to its reproductive fitness. The timing of such emergence would optimize the costs/benefits trade-off by on-setting a certain trait at a particular age, rather than earlier or later. Originally, this perspective was employed to explain the timing of morphological and physiological traits, such as sexual maturation, but recently it has been argued that it may also explain children's delay in acting accordingly to fairness principles (Sheskin, Chevallier, Lambert, & Baumard, 2014). In fact, while infants appear to evaluate others following an implicit understanding of fairness and harm, only some years later they consistently apply those moral principles during their social interactions (Siegal, 1982).

How can one apply life-history theory to the development of intent-based moral reasoning? Advocates of life-history theory may want to claim that the elicited intent-based moral reasoning emerges roughly at 5-6 years because at this time-point children increasingly engage in social interactions with peers. To understand and properly evaluate others' intentions is fundamental in forming and maintaining such relationships. Attending to agents' intentions, rather than actions outcomes, in the evaluations of agents, may become crucial just

at the age in which children, in the evolutionary past, could not rely anymore on 'free' resources provided by parents and had to rely on their interactions with peers, avoiding potentially dangerous conflicts (Marlowe, 2005). Therefore, life-history theory may explain both the growing concern for fairness and the growing reliance on agents' intention in preschoolers. The defendants of this position may then conclude that the human mind is wired with an innate ability to understand and evaluate others' intentions, but it is only during the late preschool years that this ability is systematically recruited by children in a variety of social interactions.

Conclusions

In sum, we have seen that young preschoolers' outcome-based judgment is preceded by an early capacity to evaluate intentions that is revealed in spontaneous-response tasks. Drawing a parallel with the literature on the acquisition of mental state reasoning, we argued that the outcome-to-intent shift is best explained by an expression account that posits an early emerging infant socio-moral competence and explains preschoolers' outcome-based judgments as due to immature domain-general executive function. Current evidence is more consistent with a view that assumes developmental continuity than with the opposite view based on conceptual changes.

CHAPTER 5

Mental State Understanding and Moral Judgment in Children with Autistic Spectrum Disorder

This chapter is based on the following original article:

Margoni, F., & Surian, L. (2016). *Mental state understanding and moral judgment in children with autistic spectrum disorder*. Frontiers in Psychology, 7, 1478. doi:10.3389/fpsyg.2016.01478

Mental State Understanding and Moral Judgment in Children with ASD

Do children with autistic spectrum disorder (ASD) develop the ability to take into account an agent's mental states when they are judging the morality of his or her actions? The present article aims to answer this question by reviewing recent evidence on moral reasoning on children with autism and typical development. A basic moral judgment (e.g., judgments of violations in which negative intentions are followed by negative consequences) and the ability to distinguish between conventional and moral violations appear to be spared in autism (Leslie et al., 2006). However, a closer look at the data reveals that these capacities can be explained by the tendency of ASD individuals to rely heavily on actions consequences and other external factors rather than agents' mental states. By contrast, studies that presented typically developing (TD) children with accidental and failed attempts actions have shown that even preschoolers can display an intent-based moral judgment (e.g., Cushman et al., 2013; Margoni & Surian, 2016a). The tendency to rely on outcome in ASD children is further confirmed by those studies that direcly show that ASD individuals fail to attend to the agents' intentions when the cases are more complex or ambiguous, like in accidentally harmful actions or failed attempts to harm. We propose that the impairment in understanding others' mind hinders the development of an intent-based moral judgment in children with ASD.

Mental State Reasoning in the Moral Judgment Tasks

In our social life, we often engange in the evaluation of others' actions and intentions, and we are very sensitive to harmful acts and violations of rights. For example, we maintain friedships on the basis on an assessment of our friends' moral behaviors towards us. The production and the justification of a moral judgment is a complex socio-cognitive task that often requires the use of mental state reasoning abilities (Moran et al., 2011; Young et al., 2007). In particular when people are asked to evaluate accidental harming (or helping)

actions or failed attempts to harm (or help), they need to weigh the agents' intention, that requires a mental state analysis, against the external consequences of the action. Several neuroscientific studies confirm the association between moral judgment and theory of mind (Young et al., 2010; Young et al., 2007; Young & Saxe, 2009).

Then, to what extent individuals with ASD, who present deficits in theory of mind abilities (Abell et al., 2000; Baron-Cohen et al., 1985; Baron-Cohen et al., 2000; Bowler, 1992; Castelli et al., 2002; Surian & Leslie, 1999), meet with difficulties in the acquisition of an intent-based moral judgment? Individuals with ASD are characterized by impaired social interactions and communication abilities, and a set of restricted and repetitive behaviors. Here we focus on their impairment in mentalizing, that has been shown to be a main factor affecting their socio-moral abilities. Studies on the moral judgment of ASD children have traditionally focused on a) the capacity to distinguish between moral and social-conventional transgressions and b) the ways in which individuals with autism judge the moral rightness or wrongness of an action.

Moral and Conventional Transgressions

One fundamental aspect of the moral competence has been identified by social domain theorists in the capacity to distinguish between moral and social-conventional violations. While the former involve a victim and are to be blamed regardless of the social context, the latter do not need to involve a victim and are contingent over a specific group consensus or authority mandate (Killen & Smetana, 2015; Nucci, 1981; Turiel, 1978). By the age of three, children judge moral violations, like hit someone, more harshly and less authority-dependent than social-conventional, like wearing pajamas at school (Nucci, 1985; Smetana & Braeges, 1990).

The capacity to distinguish between these two types of violation is intact in ASD individuals (Blair, 1996; Rogers et al., 2006; Shulman et al., 2012; Zalla et al., 2011). However, ASD individuals produce poorer justifications compared to TD individuals, and they do not evaluate moral violations as more serious than non-moral but disgusting actions, such as drinking tomato soup out of the bowl at a dinner party. Moreover, contrary to TD children, school-aged children with ASD are swayed by the victims' emotion and judge wrong actions that caused the crying of the victim more harshly than wrong actions that did not cause any crying (Weisberg & Leslie, 2012). ASD children usually succeed in tasks devised to investigate the moral-conventional distinction, but they rely mainly on external factors that could depend on irrelevant variables such as the particular emotional level of the agents.

The Relative Weight of Intention and Outcome in the Judgments of ASD Individuals

A working hypothesis here is that ASD children respond as TD children do when they are presented with simple, unambiguous moral cases (i.e., a negative/positive outcome produced by an intentional action with the same valence). In those cases, the difficulties encountered in integrating the mental state understanding in the moral reasoning can be overcome by the children's reliance on action outcomes and victims' emotional reactions. For this reason, ASD children appear to develop a basic moral judgment.

ASD school-aged children evaluate actions that are motivated by positive or negative intentions and are followed by congruent outcomes as TD children do (Leslie et al., 2006; Li et al., 2014). Moreover, they are able to judge an agent that caused intentionally a bad outcome more harshly than an agent that caused it accidentally, although they do not produce verbal justifications that refer to the agent's intention (Grant et al., 2005). However, Steele et al. (2003) found that children with ASD aged 4 to 14 failed to distinguish between intentional

and accidental bad acts (e.g., failing to come to a planned meeting as a result of cancelling the plan without telling or as a result of the bus breaking). Studies on ASD adults also showed that they judge an accidental harm both more punishable and more intentional compared to TD adults, suggesting a partial impairment in the ability to rely on intentions (Buon et al., 2013; see also Rogé & Mullet, 2011; Salvano-Pardieu et al., 2015; Zalla & Leboyer, 2011). However, ASD school-aged children distinguish between a distressed victim and an individual in distress that however is not a victim (Leslie et al., 2006). So, their judgments do not completely rely on the external outcomes assessment.

However, what about the judgments of more complex cases such as the failed attempts to help or harm, that require a more substantial contribution of mental state reasoning? In fact, in judging an ambiguous case such as a failed attempt to harm, it is not possible to rely solely on action outcomes and still produce a moral condemnation of the agent.

A first evidence of an outcome-bias in the judgment in ASD individuals comes from those studies that reported a 'heteronomous' (i.e., rules are understood as handed down by authority, and violations are wrong because they produce bad outcomes, namely they lead to punishment) rather than an 'autonomous' (i.e., rules are based on socially agreed-on principles, and violations are wrong because of the agent's beliefs and motivations) moral reasoning in ASD school-aged children (Grant et al., 2005; Takeda et al., 2007; see also Fadda et al., 2016). ASD children attributed moral wrongness and badness to actions that caused bad outcomes. A second and more direct evidence comes from a study that presented ASD individuals with accidental and failed attempted harms. Moran et al. (2011) found that they failed to distinguish between the two scenarios, and they judged the accidental harm significantly more harshly than TD individuals. Moreover, there is evidence of an activation of the right temporo-parietal junction (RTPJ) – an area associated with mental state reasoning – in TD individuals during the evaluation of intentional versus accidental harm, but such

result has not been found in adults with ASD (Koster-Hale et al., 2013). These results clearly suggest that ASD individuals fail to integrate the agent's mental states in their moral reasoning when judging situations in which intentions and outcomes present different valences (see Figure 1).

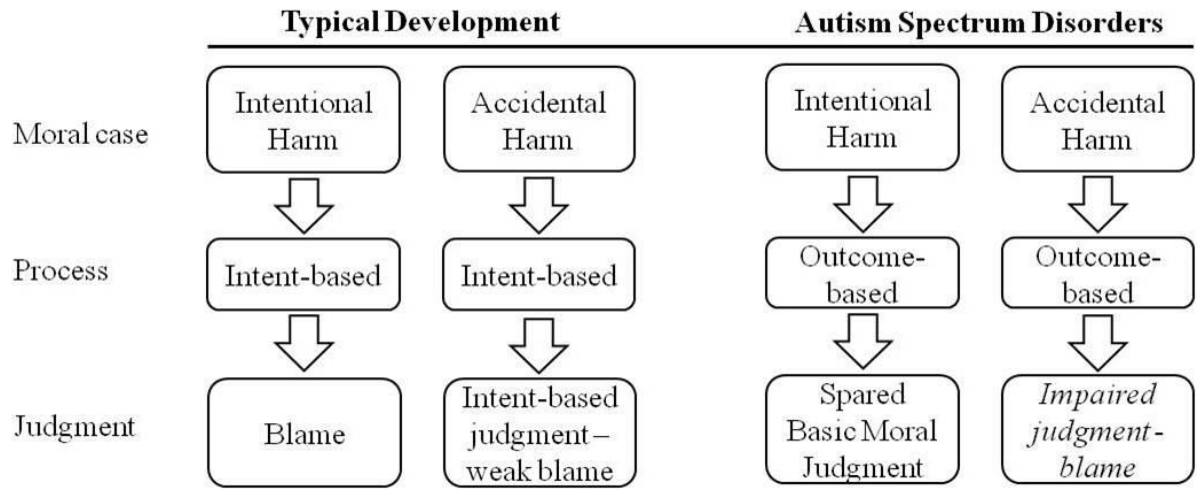

Figure 1. Main results concerning the mental state understanding in ASD individuals' moral reasoning.

Theoretical Implications of the Studies on Mental Reasoning in ASD Individuals' Moral Judgments

Three main theoretical implications relevant for the current understanding of the relationship between theory of mind and moral reasoning could be inferred from the results we briefly discussed. First, the evidence that ASD individuals, who are characterized by an impaired mental state understanding, show an atypical moral judgment, further confirms that theory of mind is fundamental for the development of a mature moral reasoning.

Second, the study of moral judgment in ASD individuals could prove useful in assessing the role of cognitive empathy in the production of a moral evaluation. ASD individuals show a spared capacity for emotional empathy (e.g., Blair, 1999; Rogers et al., 2007), that is, the

proper emotional response to others' emotions, but an impaired capacity for cognitive empathy, that is, the proper knowing how others may feel. While emotional empathy skills help ASD children developing a basic moral judgment by relying on the emotional and external aspects of the moral case such as the victims's emotional reactions or the actions outcomes (Hobson et al., 2009; Leslie et al., 2006; Weisberg & Leslie, 2012), the poor understanding of the cognitive aspects hinders the development of an intent-based moral judgment. Further studies confirm this interpretation by reporting that aspects related to cognitive empathy impairment affect the moral evaluations of ASD individuals (Channon et al., 2010; Gleichgerrcht et al., 2013; Patil et al., 2016).

A third relevant theoretical implication concerns whether the action understanding required in moral evaluation is mentalistic. A mentalistic understanding represents and explains others' actions by ascribing mental states such as beliefs, desires and internal representations to the agents (Baillargeon et al., 2010; Baron-Cohen et al., 1985; Leslie, 1987; Surian et al., 2007). By contrast, a non-mentalistic or teleological understanding represents others' actions without ascribing mental states, by linking directly the agent's actions, the goal-states and the situational constraints through the principle of rational actions (i.e., agents act to achieve certain goals choosing the most efficient means; Gergely & Csibra, 2003; Schlottmann et al., 2009). According to the proponents of teleological accounts of action understanding, humans first develop very early in life a non-mentalistic understanding, and only later they acquire a mentalistic understanding. While it could be argued that ASD individuals possess the ability to interpret actions in a non-mentalistic way already during preschool years (Hamilton, 2009; Vivanti et al., 2011), we have seen that they do not develop a mature intent-based moral judgment. Therefore, the literature on ASD individuals suggests that a non-mentalistic understanding is not sufficient for the development of a full-blown intent-based moral reasoning.

Conclusions

The ability to produce moral evaluations often requires the understanding of others' mental states and it is central for living in human social groups. While much more research is needed to acquire a full understanding of the development of moral judgment in ASD individuals, the current state of the literature suggests that this clinical population encounters some difficulties in developing a mature intent-based moral judgment because of the well-known impairment in mental state understanding. Nevertheless, ASD individuals show the ability to produce a basic moral judgment by relying on external cues such as the action outcomes and the victims' emotional reactions.

Can these results turn out to be useful in guiding programs designed to improve moral judgment in children with ASD? Since a main result of the literature we reviewed is that individuals with ASD show difficulties in integrating mental states information in their judgments, clinical treatments and educational programs aimed at improving their theory of mind abilities are likely to have, as a side-effect, a positive impact also on their moral reasoning abilities. Further research is needed to point out whether such a desiderable effect is achieved equally by any effective training on mentalizing skills (e.g., Silver & Oakes, 2001; Fisher & Happé, 2005; Begeer et al., 2011), or it is best achieved by a program that requires both mental state attribution and the generation of moral judgments.

PART 3

ADULTHOOD: INTENT-BASED MORAL REASONING

CHAPTER 6

How Intentions, Negligence and Outcomes Affect Moral Judgments

This chapter is based on the following original article:

Margoni, F., & Surian, L. (2016). *How intentions, negligence and outcomes affect moral judgments.* Manuscript submitted for publication.

Abstract

Current models of moral judgment assign a crucial role to intentions and actions outcomes in moral judgment, but they diverge on the role of negligence. In this study, a group of adults evaluated the moral rightness (or wrongness) and the deserved reward (or punishment) of a set of actions. Actions were embedded in vignettes in which agents' intentions and negligence, and actions outcomes were orthogonally manipulated. Across two experiments, we found that intention played the most important role in moral judgment. Action consequences affected more deserved reward and punishability than rightness or wrongness, and negligence played a significant but marginal role. Furthermore, we discuss the current findings in light of the pressing need to integrate existing processing models of moral judgment to account for moral approvals.

How Intentions, Negligence and Outcomes Affect Moral Judgments

When we judge the morality of an agent that helps or harms, we may rely on the agent's intention, the consequences brought about by his or her action, and the degree of his or her negligence. Very early in life, humans' socio-moral expectations are sensitive to agents' intentions (Dunfield & Kuhlmeier, 2010; Hamlin, 2013; Lee, Yun, Kim, & Song, 2015), and during preschool years our moral judgments start to rely mainly on intention, when assessed by verbal tasks (Cushman, Sheketoff, Wharton, & Carey, 2013; Karniol, 1978; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Margoni & Surian, 2016a; Piaget, 1932). Moreover, both in moral philosophy (Abelard, 1971; Kant, 1785/1959) and jurisprudence (Williams, 1953), the role of intention in judging someone responsible and morally bad or good has been highly emphasized. However, the condemnation of cases like accidental harm highlights that other factors, such as action outcomes and agent's negligence, may affect our moral evaluations (Williams, 1981).

Contrasting models of moral judgment processing have been proposed in order to account for the relative weight of intention, outcome, and negligence. A seminal model, that was proposed by attribution theorists (Heider, 1958), posited that people first analyze causal responsibility for action outcomes, and subsequently analyze internal factors such as agent's intention (Darley & Shultz, 1990; Fincham & Roberts, 1985; Shaver, 1985; Weiner, 1995). According to this model, both outcome and intention are needed to attribute responsibility, moral blame, and punishability, but causal and intentional information are integrated by a single cognitive process. For example, the punishment judgments have been described as the outcome of a process that first assesses the causal responsibility of an agent and then attributes *moral* responsibility to him or her (Shultz, Schleifer, & Altman, 1981; Shultz, Wright, & Schleifer, 1986).

A recent model, proposed by Cushman (2008, 2015; see also Young, Cushman, Hauser, & Saxe, 2007), maintains that both outcome and intention are assessed by the individual, but posits two distinct and independent cognitive processes underlying moral judgment. One process focuses on causal responsibility for harming outcomes, and another focuses on agent's mental states such as intention. This two-process model predicts conflict between processes, for example in those cases in which the agent accidentally harms someone (bad outcome without negative intention) or intends to harm someone but fails (bad intention without negative outcome). According to the model, wrongness or badness judgments rely mainly on the intent-based process, whereas punishment judgments rely both on outcome- and intent-based process. A first evidence for the independence of the two processes can be found in people judging morally bad instances of failed attempts to harm somebody, despite the fact that the agent did not cause any negative consequence (Cushman, 2008).

However, when we evaluate an accidental harming action, we may also focus on whether the agent acted with negligence. That is, we may also want to assess the manner in which the agent acted and, more specifically, decide whether he acted with or without care. We can define 'negligence' as a lack of due care in acting (e.g., Abelard, 1971; D'Arcy, 1963; Hart, 1968). Few studies examined the role of negligence in moral judgments compared to the large number of studies on intention and outcome. Some studies found that moral judgments rely on intention, but this information interacts with negligence and with the particular case being judged (e.g., Finkel & Groscup, 1997), and information about consequences affects moral judgments especially when people attribute negligence to the agent (Enzle & Hawking, 1992).

Thus, negligence could be an important factor to be accounted for by a model of moral judgment (for extensive discussion see Malle, Guglielmo, & Monroe, 2014; Weiner, 1995). Preschoolers distinguish between innocent accidental harm and accidental harm caused by

the agent's lack of due care (Schleifer, Shultz, & Lefebvre-Pinard, 1983; Siegal & Peterson, 1998). Children use negligence information to assign moral responsibility (Shultz, Wright, & Schleifer, 1986), and it has been proposed that preschoolers' judgments appear to be often outcome-based because children assume that accidental harm resulted from agent's negligence (Nobes, Panagiotaki, & Pawson, 2009).

Moreover, while negligence may not be particularly relevant in the attribution of goodness and deserved reward to agents that caused a positive outcome (e.g., help), it may instead be central in the attribution of badness and punishability for negative outcomes (e.g., harm). That is, deserved reward may be assigned only if the action was planned (intentionally), but punishment may be assigned because the action was carried out with negligence, despite the lack of any bad intention (Shultz & Wright, 1985). Therefore, particularly for harming actions, an agent can be blamed when he showed no relevant intention, and this is because he is judged a negligent agent. Adding to this, people often assume that while negative outcomes deserved a punishment, agents that caused positive outcomes are morally approvable, but they do not necessarily deserve to be rewarded. Consistently, it has been suggested that judgments of moral disapprovals rely more on attribution of causal responsibility than judgments of moral approvals (Bostyn & Roets, 2016; see also Bohner, Bless, Schwarz, & Strack, 1988).

The current research

The aim of the current study was twofold. First, we asked whether Cushman (2008, 2013)'s model has to be modified to allow for a role played by agents' negligence (Finkel & Groscup, 1997; Nobes et al., 2009; Shultz & Wright, 1985). Cushman (2008) investigated people's moral judgments using scenarios that might have been interpreted as involving some degree of negligence. However, here we make negligence information clearer as we wanted to study directly its weight in people's moral judgments. Second, we investigated whether

Cushman (2008)'s model can be generalized to judgments of moral rightness and deserved reward. Note that the computational models recently proposed to account for moral judgment have been tested mainly on the evaluation of moral transgressions. Little is currently known about how people produce judgments of goodness, praise, or deserved reward, and whether the same cognitive processes involved in judging negative cases are involved in judging positive ones. As a clear example of this bias, consider the recent claim that interpersonal harm is the "fundamental template unifying moral judgment" (Gray, Young, & Waytz, 2012).

In Experiment 1, participants evaluated the moral rightness or wrongness and the deserved reward or punishability of a set of different actions. We generated several scenarios by varying the valence and the presence of agent's intention, action outcome and agent's negligence, in order to assess the contribution of each type of information on the elicited moral judgment.

If Cushman (2008; 2013)'s model is generalizable to the evaluations of positive cases (such as helping behaviors), we should find that moral rightness and deserved reward judgments rely mainly on intention, but deserved reward judgments are also substantially affected by outcomes. However, given the recent evidence suggesting that judgments of praise rely less on causal attribution for outcomes than judgments of blame (Bostyn & Roets, 2016), we predicted that deserved reward judgment would be less outcome-based than punishment judgment. Moreover, we predicted that judgments of positive cases would be more intent-based than judgments of negative cases. Finally, if negligence plays a role in processing accidental cases, it should affect more decisions concerning agents' punishability than their deserved reward (Shultz & Wright, 1985).

Experiment 2 further investigated the role of negligence in judging the wrongness and punishability of accidental harming actions. We studied the effect of negligence (that is, the lack of care in performing an action) on participants' judgments, controlling for the

knowledge state of the agent (whether the agent was/was not aware that his behavior was dangerous), an aspect that was not always made explicit in Experiment 1 stories.

Experiment 1

Method.

Participants. Participants were 120 adults (70 female), with a mean age of 26.54 (*SD* = 4.66). Participants were recruited among students enrolled in psychology courses at the University of Trento or from the urban middle-class area surrounding the campus. All participants provided written and informed consent. The experimental procedure was approved by the local University Ethics Committee.

Materials and procedure. Participants were tested in a quiet room, at their house or in the laboratory. Each participant was invited to complete a paper-and-pen questionnaire, which took approximately twenty-five minutes. We adopted a 2 × 2 × 2 × 2 design, thus generating 16 possible and different combinations out of four factors: intention (present or absent), negligence (present or absent), consequence (present or absent), and valence of intention and outcome (positive or negative). We created four different scenario contexts, labeled “justice”, “benevolence”, “temperance”, and “help/harm”. Therefore, we created 64 different scenarios: eight combinations × four contexts (= 32 scenarios) describing *positive* events, and 8 × 4 = 32 scenarios describing *negative* events.

The contexts were inspired by three classical virtues of the Latin and Greek conceptual world and literature (e.g., see Marcus Aurelius, II century AD/2006), namely justice (*iustitia*), where the agent acts in a fairly/unfairly way, benevolence (*benevolentia* – also inspired by Cushman, unpublished data), where the agent is kind/unkind towards others, and temperance (*temperantia*), where the agent controls his anger or does not. We included a fourth context (help/harm) from the recent literature on moral psychology (Doris, 2010), where the agent specifically helps/harms another person. For each context, we generated eight positive and

eight negative different scenarios (e.g., eight temperance and eight intemperance scenarios). Below we report a single scenario context parametric variation of factors (relative to benevolence, positive scenarios only):

Background = James is seated on a subway train, when he sees a very old man standing in the corridor.
Positive intention present = James wants to be kind towards the man by giving him his own sits.
Positive intention absent = James does not have any particular intention. He simply wants to stand up in order to stretch out his legs.
Negligence present = James stands up without caring too much to be seen leaving his seat.
Negligence absent = James stands up caring to be seen leaving his seat.
Positive outcome present = James frees the seat; the man sees it and sits down.
Positive outcome absent = James frees the seat; however, the man does not see it and does not sit.

Each participant evaluated two positive and two negative scenarios for each context. Thus, participants evaluated 16 scenarios, eight positive and eight negative, presented in a randomized order. For the positive scenarios, we asked to evaluate "How morally right is [Anthony]'s behavior?" and "How much does [Anthony] deserve to be rewarded?" For the negative scenarios, participants evaluated "How morally wrong is [Anthony]'s behavior?" and "How much does [Anthony] deserve to be punished?" Each question was followed by a seven-point scale anchored at 1 with *none*, at 4 with *some*, and at 7 with *very much*. The questions order was counterbalanced between subjects.

Results and discussion.

We analyzed our data treating the mean judgment of a single scenario as the unit of analysis (as in previous and similar studies: e.g., Cushman, Young, & Hauser, 2006; Cushman, 2008). Data were averaged across the 30 participants' trials for each combination

of context and factors (30 because each one of the 120 participants evaluated only ¼ of the scenarios, that is, 16 scenario combinations out of 64).

A three-way repeated-measures ANOVA was performed, treating each context as the unit of analysis, and controlling for the questions order effect by adding to the model this variable as a between-subject factor. The repeated measures for each case were the variations in intention, negligence, and consequence. Thus, we ran our analysis on eight cases, since we had four different scenarios but two different questions orders, and we analyzed separately the judgments of negative scenarios and the judgments of positive scenarios.

The weight of intention, negligence and consequence.

Figure 1 displays the effect size of intention, negligence and consequence factors for judgments of rightness, deserved reward, wrongness, and punishment. Collapsing judgments of *moral rightness* across contexts, the proportion of the total variability accounted for the intention factor (effect size, η^2) was 72%, $F(1, 7) = 29.07$, $p = .002$, partial $\eta^2 = .83$, for the negligence was 3%, $F(1, 7) = 8.14$, $p = .029$, partial $\eta^2 = .58$, and for the consequence was lower than 1%, $F(1, 7) = 2.14$, $p = .19$, partial $\eta^2 = .26$. Intention and negligence factors contributed significantly to the model, but not consequence. Error and interactions components accounted for the remaining 24%. Interactions accounted for less than 1% of the total variability, with none of them reaching statistical significance at $p < .05$.

With respect to the *deserved reward* judgments, the proportion of the total variability accounted for intention was 52%, $F(1, 7) = 25.03$, $p = .002$, partial $\eta^2 = .81$, for consequence was 6%, $F(1, 7) = 6.24$, $p = .047$, partial $\eta^2 = .51$ and for negligence was 3%, $F(1, 7) = 10.60$, $p = .017$, partial $\eta^2 = .64$. Intention, negligence, and consequence factors contributed significantly to the model. Interactions accounted for 4% of the total variability, but none of them reached statistical significance at $p < .05$.

Thus, the intention was by far the most important factor affecting approvals, with consequence playing a minor role in determining judgments of deserved reward (6% of the total variability compared to less than 1% for rightness attribution).

For the judgments of *wrongness*, the proportion of the total variability accounted for intention was 54%, $F(1, 7) = 38.45$, $p < .001$, partial $\eta^2 = .86$, for negligence was 3%, $F(1, 7) = 19.99$, $p = .004$, partial $\eta^2 = .77$, and for consequence was also 3%, $F(1, 7) = 15.71$, $p = .007$, partial $\eta^2 = .72$. Intention, negligence, and consequence factors contributed significantly to the model. Interactions accounted for 4% of the total variability. The interaction between negligence and consequence reached statistical significance, but accounted for only 0.1% of the total variability, $p = .037$; surprisingly, participants judged more wrong an action carried out without negligence than an action carried out with negligence, especially when the action caused a negative outcome.

With respect to the *punishment* judgments, the proportion of the total variability accounted for intention was 39%, $F(1, 7) = 37.92$, $p < .001$, partial $\eta^2 = .86$, for negligence was 3%, $F(1, 7) = 18.74$, $p = .005$, partial $\eta^2 = .76$, and for consequence was 10%, $F(1, 7) = 11.48$, $p = .015$, partial $\eta^2 = .66$. Intention, negligence, and consequence factors contributed significantly to the model. Interactions accounted for 4% of the total variability. The interaction between intention and negligence reached statistical significance, and accounted for 2% of the total variability, $p = .007$; participants judged the agent without intention less punishable than the agent with a negative intention, but surprisingly this was true especially when the agent acted without the due care. In a next result section (‘Accidental actions carried out with or without negligence’)**,** we will discuss the counterintuitive interactions between intention and negligence for punishment attribution, and between negligence and consequence for wrongness attribution.

In sum, the intention was the key factor affecting disapprovals, although consequence plays also a significant role in determining punishability judgments.

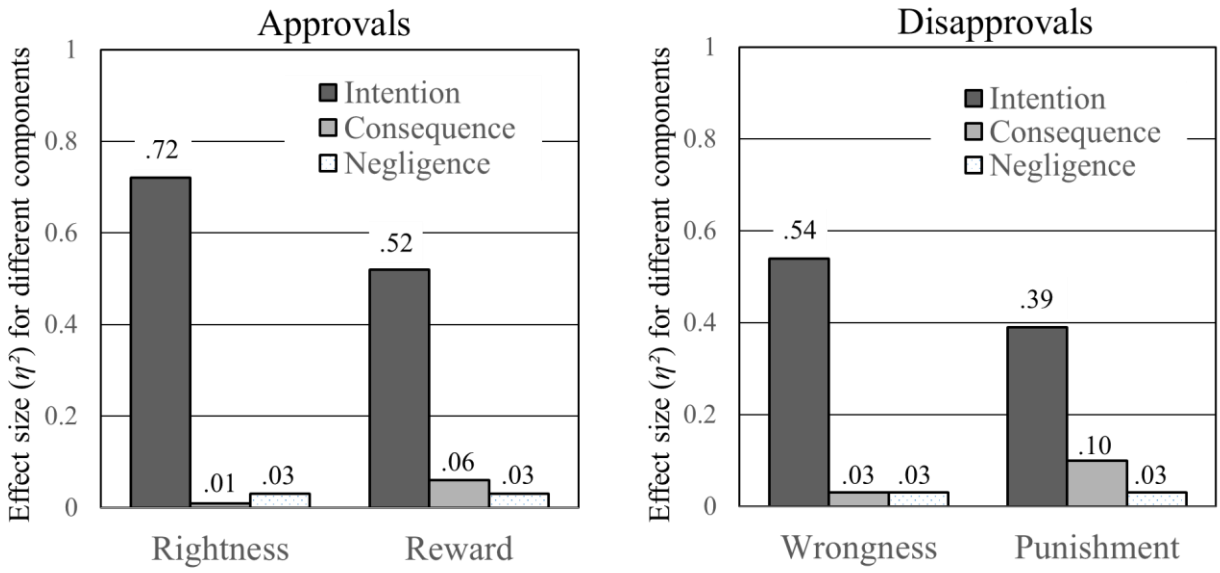

Figure 1. Proportion of the total variability explained by each factor (η^2) for the rightness and the deserved reward judgments (left), and the wrongness and the punishment judgments (right).

Moral rightness/wrongness vs. deserved reward/punishment.

In order to analyze the differences between how the factors contributed to the model for different judgments (rightness vs. deserved reward), we conducted a four-way repeated measures ANOVA, combining data sets and adding to the model a within-context factor of judgment (rightness or deserved reward). We found a significant main effect for judgment type factor (rightness vs. deserved reward), $F(1, 7) = 27.95$, $p < .001$, partial $\eta^2 = .80$. Also, the interaction between judgment type and intention was almost significant, $F(1, 7) = 5.47$, $p = .052$, partial $\eta^2 = .44$, but the interaction between judgment type and consequence did not reach statistical significance at $p < .05$.

Analyzing how factors contributed to the model for wrongness versus punishability judgment, we found a tendency toward a significant main effect for judgment type factor (wrongness vs. punishability), $F(1, 7) = 5.53$, $p = .051$, partial $\eta^2 = .44$. The interaction between judgment type and intention reached statistical significance, $F(1, 7) = 5.73$, $p = .048$, partial $\eta^2 = .45$, and the interaction between judgment type and consequence was almost significant, $F(1, 7) = 4.72$, $p = .066$, partial $\eta^2 = .40$.

Thus, our participants used intention information differently when judging rightness and deserved reward. Intention factor variability explained 72% of the total variability for rightness judgment compared to 52% for deserved reward judgment. Participants also used information about intention and, partly, consequence, differently when judging wrongness and punishability. Intention explained 54% of the total variability for wrongness judgment compared to 39% for punishability judgment, and consequence explained 3% of the total variability for wrongness judgment compared to 10% for punishability judgment. Performing a repeated-measures ANOVA that treated each participant as the unit of analysis yielded similar results.

Overall, the findings we reported above suggest that both approvals and disapprovals rely mainly on intention information. Consequence information was assessed especially in judging punishment and deserved reward. Negligence played a significant ($p < .05$) but marginal role in determining the moral judgments of participants. The current results are thus consistent with the prediction from Cushman (2008)'s model and do not fully support the request of adding negligence to the processing model of moral judgment (but see Experiment 2).

Figure 2 reports the mean judgments of rightness and deserved reward (left chart) and wrongness and deserved punishment (right chart), grouped by factor combinations of intention, negligence, and consequence. It can be easily noticed that means for deserved

reward judgment are lower or equal than means for rightness judgment, and means for deserved punishment are lower or equal than means for wrongness judgment. We asked whether varying a factor value level determines a change in one judgment type (e.g., rightness) that is different from the change determined in the other judgment type (e.g., deserved reward). We performed several two-way repeated-measures ANOVA with judgment type (rightness vs. deserved reward or wrongness vs. deserved punishment) and information (intention present vs. absent or negligence present vs. absent or consequence present vs. absent) as factors.

Positive cases. We found a significant condition-by-intention interaction, $F(1, 7) = 16.27$, $p = .005$, partial $\eta^2 = .70$; when the agent caused a positive outcome without negligence, rightness judgment average decreased significantly more than deserved reward did by the fact that the agent had no positive intention while acting. We also found a significant condition-by-negligence interaction, $F(1, 7) = 9.55$, $p = .018$, partial $\eta^2 = .58$; when there was no positive intention or outcome occurring, rightness judgment average decreased significantly more than deserved reward did by the fact that the agent was negligent. These data suggest that moral rightness judgments rely more on intention and negligence than deserved reward judgments. Finally, we found a significant condition-by-consequence interaction, $F(1, 7) = 7.44$, $p = .029$, partial $\eta^2 = .51$; when there was no intention or negligence, deserved reward judgment average increased significantly more than rightness judgment average did by the fact that the agent accidentally caused a positive outcome. Consistently with a generalization from Cushman (2008)'s model, the last interaction further suggests that judgments of deserved reward are more outcome-based than judgments of moral rightness.

Negative cases. When the agent acted with due care and did not cause any negative outcome, wrongness judgment increased significantly more than deserved punishment

judgment as a result of including information about the agent's negative intention, $F(1, 7) = 5.26$, $p = .056$, partial $\eta^2 = .43$. Moreover, when the agent had a negative intention and acted negligently, deserved punishment judgment average decreased significantly more than wrongness judgment average did by the fact that the agent did not cause the negative outcome, $F(1, 7) = 8.38$, $p = .023$, partial $\eta^2 = .70$. Taken together, these interactions replicate previous results from Cushman (2008) and show that judgments of moral wrongness rely more on intention compared to judgments of punishability that are more outcome-based.

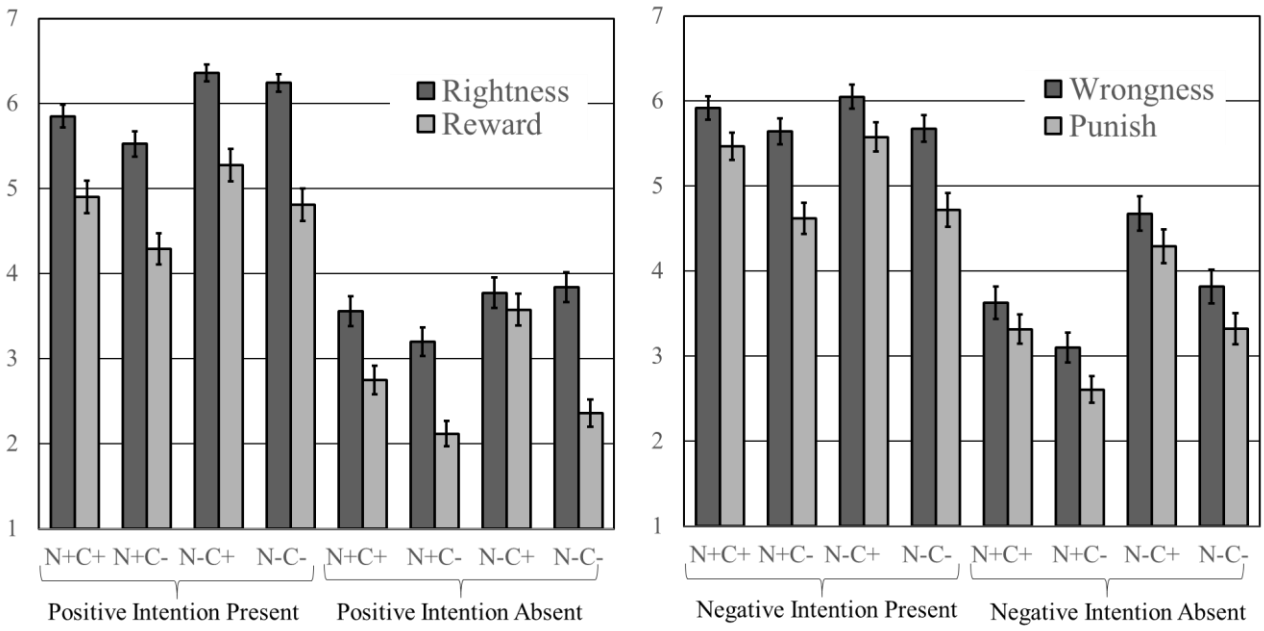

Figure 2. Mean judgment of rightness and deserved reward concerning positive actions (e.g., help; left) or wrongness and punishment concerning negative actions (e.g., harm; right) as a function of agents' intention and negligence, and actions consequence. Negligence: N+ = careless agent; N- = careful agent; Action consequence: C+ = information about the positive (left) or negative (right) outcome included; C- = scenarios made explicit that the actions had neither positive nor negative outcomes. Error bars show the magnitude of the standard error.

Accidental actions carried out with or without negligence.

One result concerning the processing of negligence information in cases of accidental harm is particularly striking. We found that a caring action that causes accidentally a bad outcome was judged more wrong, $t(119) = 4.23$, $p < .001$, and punishable, $t(119) = 4.07$, $p < .001$, than a careless action that brings accidentally a bad outcome. This is a counterintuitive result, since we ordinarily think that a careless driver that kills accidentally someone is more responsible than a careful driver that kills accidentally.

A first possibility is that we attend to negligence information only in those cases where the outcome is a severe harm to someone or something (e.g., the road accident). To test this hypothesis, we analyzed separately scenarios with a severe harm (temperance and help/harm) and scenarios with a minor harm (justice and benevolence). If people attend to negligence only when evaluating severe accidental outcomes, we should find that an accidental outcome is deemed more wrong and punishable than a negligent accidental outcome only when the harm is minor. However, this was not the case. We found the same pattern described above both when participants evaluated minor and severe harm. An accidental minor outcome was deemed more wrong, $t(59) = 3.60$, $p < .001$, and punishable, $t(59) = 3.60$, $p < .001$, than a negligent accidental minor outcome; and an accidental severe outcome was deemed more wrong, $t(59) = 2.43$, $p = .018$, and punishable, $t(59) = 2.20$, $p = .032$, than a negligent accidental severe outcome (Fig. 3).

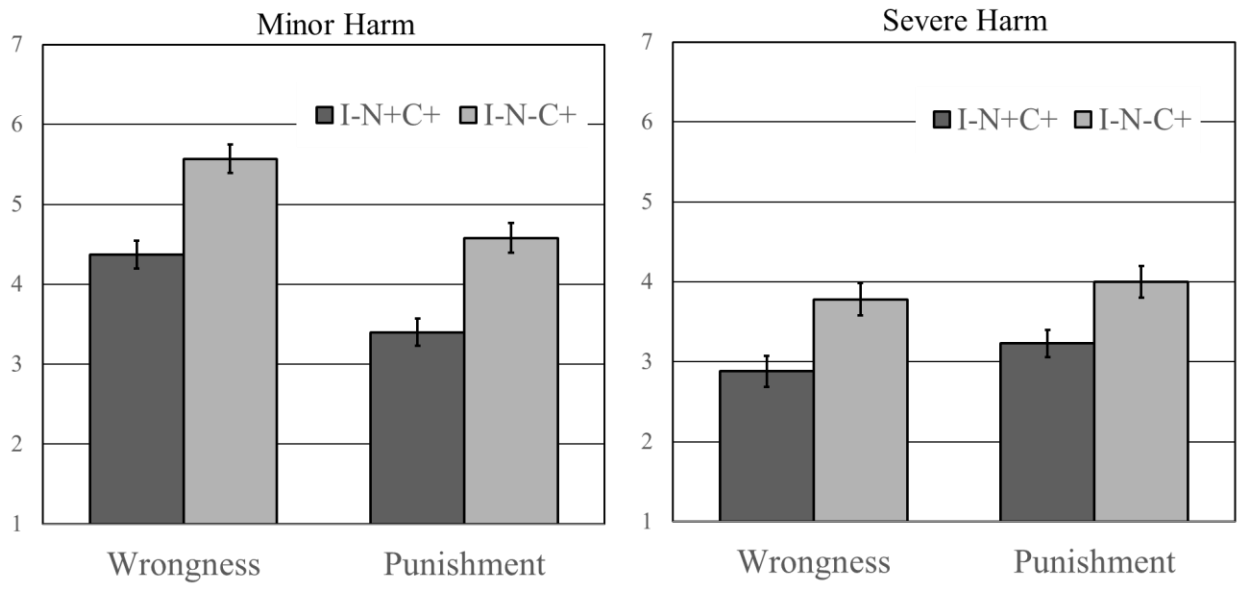

Figure 3. Mean judgment across scenario contexts with minor harm (justice and benevolence, left), or severe harm (temperance and help/harm, right) grouped by two combinations of intention, negligence, and consequence ($I^{-}N^{+}C^{+}$ = negligent accidental harm; $I^{-}N^{-}C^{+}$ = accidental harm). Error bars show the magnitude of the standard error.

A second possible explanation for the results displayed in Figure 3 could be that participants inferred a negative intention from the fact that the action was performed carefully, and perhaps they inferred from the fact that the character acted carelessly that he did not have any clear intention to harm. Then, acting with care may have been perceived as a clue of a hidden and relevant bad intention, whereas acting without care may have conveyed more clearly to our participants the absence of any evil intention. Since intention matters the most in evaluating our scenarios (see Fig. 1), negligent accident harm (less intention inferred) was judged less harshly than accidental harm (more intention inferred).

If this hypothesis is true, we should integrate Cushman (2008; 2015)'s model positing that people evaluate intention as the source of control over the following action. Negligence information could be added in the model as a signal of a weak or strong control link between

intention and action. Inferring a weak control link will eventually result in attributing less (negative) intention to the agent who harmed, and will lead to a lenient moral evaluation. By contrast, inferring a strong control link may result in attributing a negative intention and evaluating the agent harshly because of his alleged evil intention.

Experiment 2

The results of Experiment 1, showing that accidental harm caused by a careful agent is condemned more than a negligent accidental harm, may be due, at least in part, to processing demands of the task, which required the evaluation of 16 different scenarios. In this demanding context, it is possible that information about agents' negligence have been misused by participants as a cue to infer agents' 'true intentions'. To test this hypothesis, we conducted a second experiment in which participants evaluated a minor number of scenarios (n = 4). Moreover, we noticed that in Exp. 1 stories, the carefulness with which the agent acted (that is, the negligence) and the agent's knowledge state (that is, whether the agent was aware of the dangerousness of the action) were sometimes confounded. Because the strength of the control link between intention and action may be inferred by both the carefulness with which the agent acted and the agent's knowledge state, in Experiment 2 we investigated the role of negligence (lack of due care in acting) in judging accidental harm cases controlling for the agent's knowledge state information. Participants rated with a 7-point-scale the moral wrongness and the punishability of a minor number of accidental harming actions that are carried out by the agent either with or without care and relevant knowledge.

Method.

Participants. Participants were 48 adults (36 female, Mean age = 21.67, *SD* = 1.65), recruited through their courses at the Department of Psychology, University of X. All participants provided written and informed consent.

Materials and procedure. Participants were tested in groups with the supervision of one experimenter. Each participant was invited to complete a brief paper-and-pen questionnaire. We adopted a 2 × 2 design, with two story contexts. We generated eight accidental harm stories in total. The two factors were *care* (the agent acted with/without the due care) and *knowledge state* (the agent was/was not aware of the dangerousness of her action). For each context, we generated four stories: two stories were the agent was informed and either acted with or without negligence; and two stories were the agent was not informed and acted with or without due care. A first context was about a waitress that accidentally tramples on a cradle containing a baby; whether she acted carefully and knew the baby's position varied. A second context was about a man that took his child to a mountain hike and accidentally knocks him into a ravine; whether he moved carefully and heard the weather forecast varied.

Each participant judged two scenarios taken from the 'waitress' context and two from the 'mountain' context. Participants evaluated with a seven-point scale (1 = none; 4 = some; 7 = very much) "How morally wrong is [character's name]'s behavior?", "How much does [character's name] deserve to be punish?", and "How much responsibility does [character's name] have for what happened?". We added a responsibility question in order to control whether any difference in the evaluation of wrongness or punishability between cases could be related to a difference in the attribution of responsibility for the outcome. Scenarios and questions orders were counterbalanced between participants using a Latin Square.

Results and discussion.

We performed a two-way MANOVA, with care and knowledge state as independent variables and wrongness, punishability, and responsibility judgment as dependent variables. We found a main effect of care, $F(1, 186) = 11.47$, $p < .001$, Wilks' $\Lambda = .84$, and a main effect of knowledge, $F(1, 186) = 10.96$, $p < .001$, Wilks' $\Lambda = .85$, but no significant

interaction between them. For wrongness, punishability, and responsibility judgments there were an effect of care (wrongness, $F(1, 191) = 16.55$, $p < .001$, partial $\eta^2 = .08$; punishability, $F(1, 191) = 12.16$, $p < .001$, partial $\eta^2 = .06$; responsibility, $F(1, 191) = 31.98$, $p < .001$, partial $\eta^2 = .15$) and an effect of knowledge state (wrongness, $F(1, 191) = 6.23$, $p = .013$, partial $\eta^2 = .03$; punishability, $F(1, 191) = 8.98$, $p = .003$, partial $\eta^2 = .05$; responsibility, $F(1, 191) = 33.06$, $p < .001$, partial $\eta^2 = .15$).

With a series of t-test, we further analyzed the results pattern. When comparing the accidental harm caused by an agent acting with care and informed (C^+K^+) with the accidental harm caused by a careless but informed agent (C^-K^+), all comparisons were significant and C^+K^+ was judged always leniently (Fig. 4). C^+K^+ was judged less wrong ($M = 2.15$, $SD = 1.5$) compared to C^-K^+ ($M = 3.13$, $SD = 1.87$), $t(94) = 2.83$, $p = .006$. Participants attributed less responsibility for C^+K^+ ($M = 3.85$, $SD = 1.77$) compared to C^-K^+ ($M = 5.21$; $SD = 1.17$), $t(94) = 4.44$, $p < .001$. They also judged C^+K^+ less punishable ($M = 2.31$, $SD = 1.34$) compared to C^-K^+ ($M = 3.27$, $SD = 1.65$), $t(94) = 3.13$, $p = .002$. Also when the agent was ignorant, comparing the caring case (C^+K^-) with the careless one (C^-K^-) led to similar results. Participants blamed more the agent acting without due care.

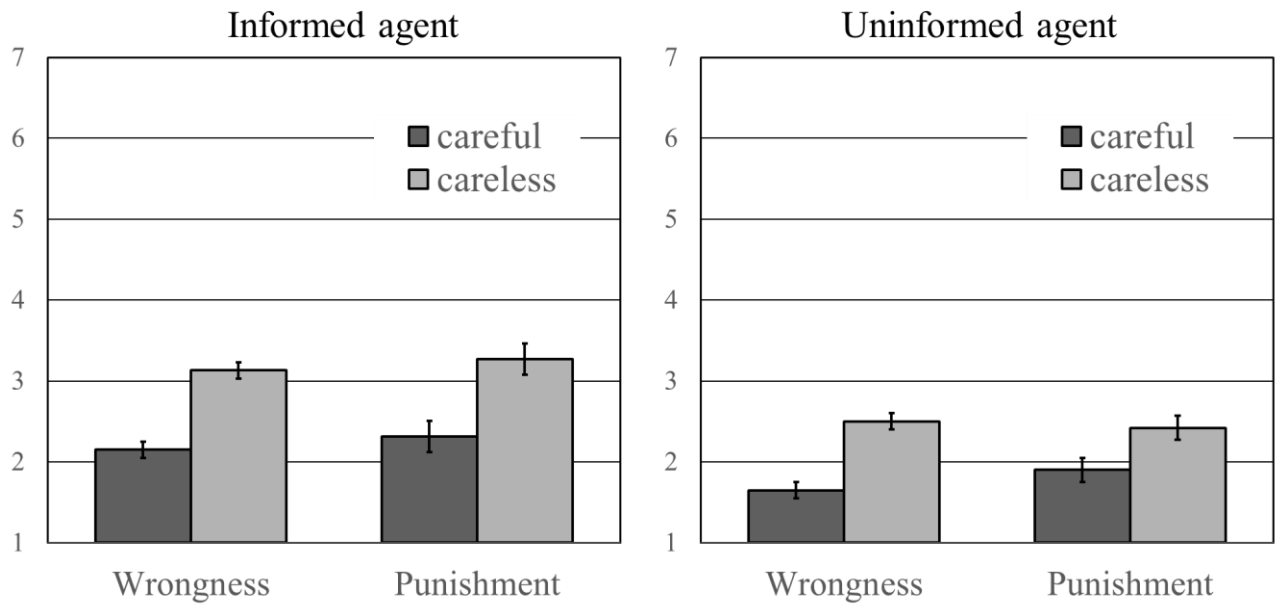

Figure 4. Mean wrongness and punishability judgments of the accidental harm stories in which the agent was informed (left) or uniformed (right), from Experiment 2. Error bars show the magnitude of the standard error.

In sum, we found that in some cases negligence information is taken into account, and that, in this cases, people judge accidental actions carried out with negligence more wrong and punishable than accidental actions carried out without negligence. Results from Experiment 2 suggest that the underestimation of negligence information and the counterintuitive result of Experiment 1 about the use of negligence in judging accidental harm cases may be linked with the task processing demands and the less explicit information about the agents' knowledge state. When people have to evaluate 16 scenarios, and information about negligence may be taken as a cue to agents' intentions, then the participants' attention is likely to shift towards intentions and outcomes.

General Discussion

In Experiment 1, we investigated the relative weight of intention, negligence and outcome information in moral approvals (judgments of rightness and deserved reward) and moral disapprovals (judgments of wrongness and punishability). We found that moral judgments rely mainly on intention, and that deserved reward and punishment judgments are respectively more outcome-based than judgments of moral rightness or wrongness. We also found that negligence information play a marginal role in determining the moral judgments. In Experiment 2, we further investigated the role of negligence in judging the moral wrongness and punishability of accidental harming actions. In these cases, negligent agents were blamed reliably more than careful ones.

The major role of intention.

The results from Experiment 1 showed that people rely mainly on intention when approving or disapproving others' actions: "Ethically, intention is everything" (Piaget, 1932; p. 328). Furthermore, our results showed that moral judgments are scarcely affected by the assessment of the agent's negligence. Therefore, the current study provides evidence for a processing model of moral judgment that relies mainly on intention and outcome (e.g., Cushman, 2008; 2013). However, the results of Experiment 2, together with previous findings (e.g., Finkel & Groscup, 1997; Nobes et al., 2009; Shultz & Wright, 1985), invite a minor revision of the model in order to account for people's ability to attend to agents' negligence at least when judging accidental harm. In this sense, negligence information may not be difficult to evaluate, as Experiment 2 showed, but sometimes people may have difficulties to detect negligence information, as Experiment 1 suggested.

This conclusion is consistent with the results from Nobes et al. (2009) showing that preschoolers already rely on negligence information in judging moral cases. While it is true that our results suggest that negligence play a marginal role in determining people's moral

judgments, in Experiment 2 we find that when judging accidental harm cases (that is, the cases judged by children in Nobes et al.) people rely also on negligence information. However, Nobes et al. (2009) also argued that older children rely less on negligence compared to younger children. Therefore, while negligence may be relevant in explaining the outcome-based judgment of younger preschoolers (they condemn accidental wrongdoers by assuming they were negligent), it may still be marginal in describing the adults' intent-based moral judgments, except when adults judge accidental harming actions.

Indeed, by claiming that intention is everything, we usually imply that mental states overall are fundamental to assess the morality of an action, and negligence too can be conceptualized as a part of the agent's mental states. Although intention may be often conflated with negligence in real-life cases, in our experiments we manipulated this information in order to assess their relative weight. What our data made clear is that intention is probably the more relevant cue to assess morality among the mental states cues, and negligence is taken into account especially in judging accidental cases.

The generalizability from disapprovals to approvals.

For the first time, we asked whether the model proposed by Cushman (2008) could be generalized also to the moral approvals of positive actions such as helping. In addition to having replicated previous findings showing that moral wrongness judgments are intent-based and punishment judgments are more outcome-based compared to wrongness judgments (Cushman, 2008; Young et al., 2007), here we reported evidence of the model generalizability to approval judgments. First, as predicted, both judgments of moral rightness and deserved reward relied mainly on intention information. Second, we found the predicted difference between judgments of rightness and judgments of deserved reward, that is, moral rightness judgments were more intent-based than deserved reward judgments, and the latter were more outcome-based than the former.

However, our study also detected some differences between approvals and disapprovals that needed to be taken into account while revising the model proposed by Cushman (2008). Our study provides only a first glance on these differences, and further studies should directly address them by using positive and negative scenarios that could be directly compared. As predicted by previous results reporting asymmetries between praise and blame (Bostyn & Roets, 2016), we found that deserved reward judgments were less outcome-based than punishment judgments. Consistently with Shultz & Wright (1985) but somehow at odds with results suggesting that harmful side effects are judged as intentionally caused more often than helpful side effects are (e.g., Knobe, 2003; but see Haupt & Uske, 2012), our results suggest that intention is particularly central in assessing the moral rightness and the deserved reward rather than the moral wrongness and the punishability. By contrast, action outcomes and, sometimes, agent's negligence may be used more in assessing moral wrongness and punishability.

Future studies should be devoted to further address the issue of whether existing processing models of moral judgment could be generalized to account also for approvals of actions and agents. Moreover, within this future line of research, particular attention could be devoted to test whether intention, outcome and negligence play a different role in the evaluation of actions that simply fulfill existing norms or duties versus supererogatory actions that go beyond the call of duty.

Concluding remarks.

In sum, we reported evidence consistent with current models predicting that moral judgments are mainly intent-based, that attributions of punishment and deserved reward are outcome-based, and that negligence information is taken into account especially when judging accidental cases, but overall it plays a marginal role. Moreover, we started to clarify

in which way current models may be integrated to account also for the moral judgments of positive situations.

While negligence appears to play a marginal role in people's moral judgments, it has a crucial and widespread role in causing serious damages, for example on the road or in work environments (e.g., Poama, 2012; Reamer & Racette, 2015). The results of the current study suggest that such a role is likely to be underestimated by common sense. Therefore, our findings have potential far-reaching practical implications. In fact, they highlight the need to take these aspects of moral judgment into account when planning and carrying out effective intervention programs aimed at reducing risk in daily life activities.

CHAPTER 7

Moral Judgment in Old Age: Evidence of an Intent-to-Outcome Shift

This chapter is based on the following original article:

Margoni, F., Geipel, J., Surian, L., & Hadjichristidis, K. (2016). *Moral Judgment in old age: Evidence for an intent-to-outcome shift.* Manuscript submitted for publication.

Abstract

We examined whether aging influences the extent to which people weigh an agent's intention and the outcomes of his or her action in moral evaluations. We presented older (63–90 years) and younger adults (21–39 years) with a series of scenarios illustrating either harmful or helpful actions. Each scenario described the intention of an agent (neutral vs. harmful/helpful) and the outcome of his or her action (neutral vs. harmful/helpful). Participants had to rate how morally good or bad was the agent's action. We found that older, as opposed to younger, participants relied less on intentions and more on outcomes, but mainly in the evaluation of harmful actions. Importantly, this age-related difference was associated with older adults' decline in theory of mind abilities. We discuss theoretical and practical implications.

Moral Judgment in Old Age: Evidence of an Intent-to-Outcome Shift

The United Nations considers as older adults individuals aged 60 years or over (United Nation, 2012). In 2015, the percentage of older adults was 12% and this figure is expected to double by 2050 (World Health Organization, WHO, 2016). As the world population is aging, many people extend their careers into their golden years. For example, about 12% of 1,200 sitting federal district and circuit judges in the US are 80 years or older (Goldstein, 2011), while the average age of members in the US Senate is 61 (Manning, 2015). Thus, many highly consequential decisions that involve moral issues are taken by older adults. However, existing theories of adults' moral judgment and decision making are based on research utilizing predominantly university students (e.g., Henrich, Heine, & Norenzayan, 2010). The main aim of the present article is to examine whether and how aging affects moral judgment. If it does, we would need to re-examine our moral judgment theories.

One crucial component of moral judgment involves the consideration of mental state information such as intentions, beliefs, and desires (for a review see Young & Tsoi, 2013). For example, people typically judge *intentionally* harmful acts (e.g., intentionally poisoning someone) as morally worse than *accidentally* harmful acts (e.g., accidentally poisoning someone), although both actions may result in the same outcome. This is known as intent-based moral judgment and develops by the age of 5 to 6 (e.g., Cushman, Sheketoff, Wharton, & Carey, 2013; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Margoni & Surian, 2017). Prior to that age, moral judgment elicited through verbal descriptions is predominantly outcome-based. This outcome-to-intent-shift is associated with changes in theory of mind abilities and executive functioning skills (Chandler, Sokol, & Hallett, 2001; Killen et al., 2011; Zelazo, Helwig, & Lau, 1996).

Lifespan research, on the other hand, suggests that aging correlates with a decrease in theory of mind skills (for a meta-analysis see Henry, Phillips, Ruffman, & Bailey, 2013) and

general cognitive abilities, such as executive functions, working memory capacity, and processing speed (e.g., Amieva, Phillips, & Della Sala, 2003; Maylor, Moulson, Muncer, & Taylor, 2002; Moran, 2013; Salthouse, 2004; Sullivan & Ruffman, 2004, for a review see Reuter-Lorenz & Sylvester, 2005). Merging evidence from moral judgment and life-span research, we predicted that older adults, as opposed to younger adults, will be less likely to make intent-based and more likely to make outcome-based moral judgments. Furthermore, we predicted that these age differences would be related to older adults comparatively diminished theory of mind abilities and executive functioning skills.

Indirect support for our claim comes from a longitudinal study which found that moral reasoning stage (Kohlberg, 1969, 1984) increases sequentially throughout early development but decreases during old adulthood (Armon & Dawson, 1997; but see also Pratt, Diessner, Pratt, Hunsberger, & Pancer, 1996). Direct support for our claim comes from the work by Moran and colleagues (Moran, Jolly, & Mitchell, 2012). Moran et al. (2012) asked 14 older participants and 27 younger participants to make a series of moral judgments concerning harmful actions. Specifically, they presented participants with moral scenarios containing information about an agent's intention (neutral vs. harmful) and the outcome of his or her action (neutral vs. harmful). Participants performed the moral judgment task inside an MRI scanner and under time pressure. The authors found that older, as compared to younger, participants relied relatively less on the agent's intentions than on the actions outcomes when judging the permissibility of harmful acts. Using functional magnetic resonance imaging (fMRI), the authors further found that this effect was associated with age-related impairments in the dorsal sub-region of the medial-prefrontal cortex (MPFC), a brain region related to social cognition such as mental state reasoning. In light of their findings, Moran et al. (2012) suggested that aging effects in moral judgment may be related to an impairment of theory of

mind abilities but also to a more general cognitive decline (e.g., executive functioning). Here, we tested these possibilities empirically.

The aim of the present study was to consolidate and extend the behavioural findings of Moran et al. (2012), as well as to examine the underpinning mechanisms of aging effects on moral reasoning. To this end, we tested a greater number of older and younger adults than these authors did, in a more naturalistic context, and without imposing time constrains. In terms of materials, along with scenarios involving harmful actions we also tested scenarios involving helpful actions. To examine the underlying mechanisms of eventual aging effects on moral judgment, we included a theory of mind, a general cognitive ability, and an executive function tasks. We also included an empathic concern task because empathic concern has been shown to influence moral judgment (e.g., Choe & Min, 2011; Crockett et al., 2010; Decety, & Cowell, 2014; Gleichgerrcht, & Young, 2013; Kahane, Everett, Earp, Farias, Savulescu, 2015; Patil & Silani, 2014) and to increase with age (e.g., Sze, Gyurak, Goodkind, & Levenson, 2012).

Methods

Participants.

Thirty younger adults (20 female, M_{age} = 29.4, age range: 21—39 years)[1] were recruited through flyers posted at the campus of the University of Trento. Thirty older adults (24 female, M_{age} = 77.5, age range: 63—90 years) were recruited through a local Association for older adults. All participants took part in the study on a voluntary basis. On average participants indicated that they had 11.72 years of school education ($M_{older\ adults}$ = 8.80 years, $M_{younger\ adults}$ = 14.63 years). The University of Trento Ethics Committee approved the research protocol of the present study.

Materials and Procedure.

The experiment was conducted in a single session that lasted about 60 minutes. Participants were asked to complete a moral judgment task followed by a battery of tasks measuring individual differences.

Moral judgment task. Each participant received eight scenarios (adapted from Young, Scholz, & Saxe, 2011). Four scenarios involved ultimately harmful actions (harm scenarios) and four ultimately helpful actions (help scenarios). Within each type of scenario, we varied orthogonally the nature of the agent's intention (neutral vs. valenced) and the outcome of his or her action (neutral vs. valenced), resulting in four different trials for harm scenarios and four for help scenarios: neutral-intention/neutral-outcome, neutral-intention/valenced-outcome, valenced-intention/neutral-outcome, and valenced-intention/valenced-outcome. Table 1 presents the four trials of a harm scenario and those of a help scenario. Notice that for harm scenarios 'valenced' refers to 'negative' or 'harmful', whereas for help scenarios 'valenced' refers to 'helpful' or 'positive'.

Table 1.

Schematic Representation of the Four Different Versions of a Harm and a Help Scenario.

Background	
Harm scenario	Help scenario
Simon is grocery shopping for his grandmother who adores spinach. Recently there had been bacterial contamination of bagged spinach. At the market, Simon sees some bagged spinach on sale.	Anne is doing some shopping at the mall, when she sees many lovely bracelets to choose from. One of the bracelets has very beautiful pink stones.
Negative Intention–Negative Outcome	Positive Intention–Positive Outcome
He thinks that bagged spinach may still be contaminated because of an incident just that day in his town. Simon buys, even though he thinks it may be dangerous, his grandmother the spinach, and she cooks it ending up in the hospital, violently ill.	Browsing the counter, Anne thinks that the proceeds from the sale of the pink bracelet will contribute to breast cancer research. The sales profits from this bracelet will go directly to the cancer clinic in the city to fund breast cancer research. Anne buys the bracelet. Anne's money is used to fund breast cancer research.
Negative Intention–Neutral Outcome	Positive Intention–Neutral Outcome
He thinks that bagged spinach may still be contaminated because of an incident just that day in his town. However, contrarily to what Simon thinks it is safe to eat spinach because it is no longer contaminated, in fact bagged spinach has been restocked at many markets. Simon buys even though he thinks it may be dangerous his grandmother the spinach, and she cooks it. However, the meal is healthy and delicious.	Browsing the counter, Anne thinks that the proceeds from the sale of the pink bracelet will contribute to breast cancer research. The sales profits from this bracelet will go directly to the jewelry company to fund a new advertising campaign. Anne buys the bracelet. Anne's money is used by the jewelry company to fund advertising.
Neutral Intention–Negative Outcome	Neutral Intention–Positive Outcome
He thinks that it's perfectly safe now because someone told him so. Bagged spinach has been restocked at many markets, but some inspections aren't thorough and contaminated batches are missed. Simon thinking that it is not dangerous buys his grandmother the spinach, and she cooks it, ending up in the hospital, violently ill.	Browsing the counter, Anne thinks that the proceeds from the sale of the pink bracelet will contribute to a company profits. However, without her knowledge, the sales profits from this bracelet will go directly to the cancer clinic in the city to fund breast cancer research. Anne buys the bracelet. Anne's money is used to fund breast cancer research.
Neutral Intention–Neutral Outcome	Neutral Intention–Neutral Outcome
He thinks that it's perfectly safe now because someone told him so. It is safe to eat spinach because it is no longer contaminated, in fact bagged spinach has been restocked at many markets. Simon buys his grandmother the spinach, and she cooks it. The meal is healthy and delicious.	Browsing the counter, Anne thinks that the proceeds from the sale of the pink bracelet will contribute to a company profits. The sales profits from this bracelet will go directly to the jewelry company to fund a new advertising campaign. Anne buys the bracelet. Anne's money is used by the jewelry company to fund advertising.

Note. For all four versions of a given scenario, the background information was the same.

Following each scenario, participants were asked to judge the moral badness (for harm scenarios) or the moral goodness (for help scenarios) of the described action ("How morally bad/good was the [agent's action]?") on a scale ranging from 0 (*not at all*) to 10 (*very much*). Next, participants were asked "How much punishment/reward does the [agent's action] deserve?" and answered on a scale that ranged from 0 (*not at all*) to 10 (*very much*). For the sake of brevity, we omit the results of the punishment/reward judgments in this paper. Note that the pattern of the results was similar to that of the badness/goodness judgments (for details see Supplemental Material at the end of this chapter). The order of scenarios (harm first vs. help first) and test questions (bad/good first vs. punishment/reward first) were counterbalanced across participants.

Individual differences tasks. Following the moral judgment task, participants completed four tasks: theory of mind, empathic concern, cognitive ability level, and executive function.

Theory of mind. Participants received the Reading the Mind in the Eyes test (RME, Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Serafin & Surian, 2004). In RME, participants have to choose which one among four words best describes the mental or emotional state of a person on the basis of a picture of his or her eye-gaze. Participants were presented with 36 different pictures, and subsequently made 36 choices.

Empathic concern. Participants were presented with the empathic concern subscale of the Interpersonal Reactivity Index (IRI-EC; Davis, 1980). This subscale consists of seven items that are rated on a 5-point scale, which ranges from 1 (*does not describe me well*) to 5 (*describes me well*). This subscale assesses participants' self-reported feelings of sympathy and concern for unfortunate others.

Cognitive ability. We asked participants to complete the Digit Symbol Substitution Test (DSST of the WAIS; Wechsler, 1981). Participants were asked to complete as many items as possible within 90 seconds. This test consists of a code table displaying nine different pairs of digits and symbols. The rows of the table consist of 94 double boxes with a digit and a white space next to it. Participants are asked to fill the white space next to each digit with the appropriate symbol based on the code table. We used the DSST as a measure of general cognitive ability.

Executive functions. Participants also received the Wisconsin Card Sorting Test (WCST; Heaton, 1995). This test measures participants' executive functioning skills. Participants are asked to sort cards containing colored geometric forms of different shapes and numbers to 4 target cards. Participants are informed whether each sort is correct or incorrect. Once a participant has reached a certain number of correct sorts, the rules are changed and the participant must apply the new rule. There are a number of different test scores which can be computed. In the subsequent analyses, we focused on the number of perseverative responses. That is the number of incorrect responses that would have been correct for the preceding rule.

Results

We first examined whether any of our older participants suffered from dementia by scrutinizing their scores on the Mini-Mental Status Examination (MMSE, Folstein et al., 1975). MMSE scores between 30–24 reveal no impairment, 24–20 suspected impairment, 19–17 mild impairment, 16–10 moderate impairment, and 9–0 severe impairment/dementia. None of the older participants showed a significant age-related decrement in cognitive performance (MMSE scores were higher than 20), although three participants showed a suspected impairment (MMSE scores between 24–20). Excluding these participants from the

data analyses had no effect on the main pattern of the results. Below we report the analyses on the full sample.

Moral Judgments.

Based on previous literature that supports an intent-to-outcome shift in older adults, we predicted two interactive effects: one between age and intention, and another between age and outcome. Specifically, we predicted that older versus younger participants would weigh less intentions and more outcomes in their moral evaluations. We analyzed moral judgment with a 2 (Age: Old vs. Young) × 2 (Intention: Neutral vs. Valenced) × 2 (Outcome: Neutral vs. Valenced) × 2 (Context: Harmful vs. Helpful) mixed-factor analysis of variance (ANOVA), with age as a between-participants factor and all other factors as repeated measures. The analysis revealed a main effect of intention, $F(1, 54) = 171.55$, $p < .001$, $f = 1.78$, which was qualified by a significant Age × Intention interaction, $F(1, 54) = 23.85$, $p < .001$, $f = .66$. As expected, the judgments of older participants were less affected by agent's intention than those of younger participants (see Figure 1).

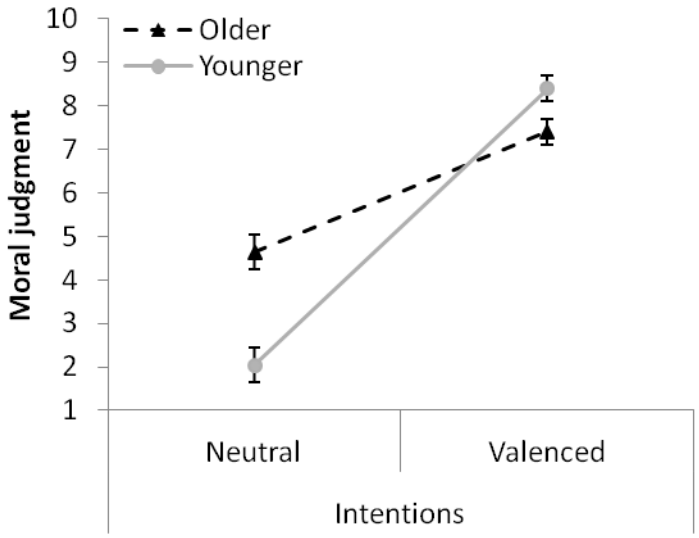

Figure 1. Moral judgment ratings by age (older vs. younger) and intentions (neutral vs. valenced). Older versus younger participants were less affected by intention status. Error bars indicate standard error of the mean.

The Age × Intention interaction, in turn, was qualified by a three-way Age × Intention × Context interaction, $F(1, 54) = 9.94$, $p = .003$, $f = .42$. We scrutinized this interaction with two separate 2 (Age) × 2 (Intention) × 2 (Outcome) analyses of variance, one for each context. The analysis for harmful contexts revealed a significant Age × Intention interaction, $F(1, 56) = 58.08$, $p < .001$, $f = .87$, with older participants being less affected than younger participants by the agent's (harmful) intention (see Figure 2). A similar analysis for helpful contexts revealed no significant Age × Intention interaction, $F(1, 56) = 2.88$, $p = .098$, $f = .23$. In sum, older participants were less sensitive than younger participants to intentions, but only in harmful contexts.

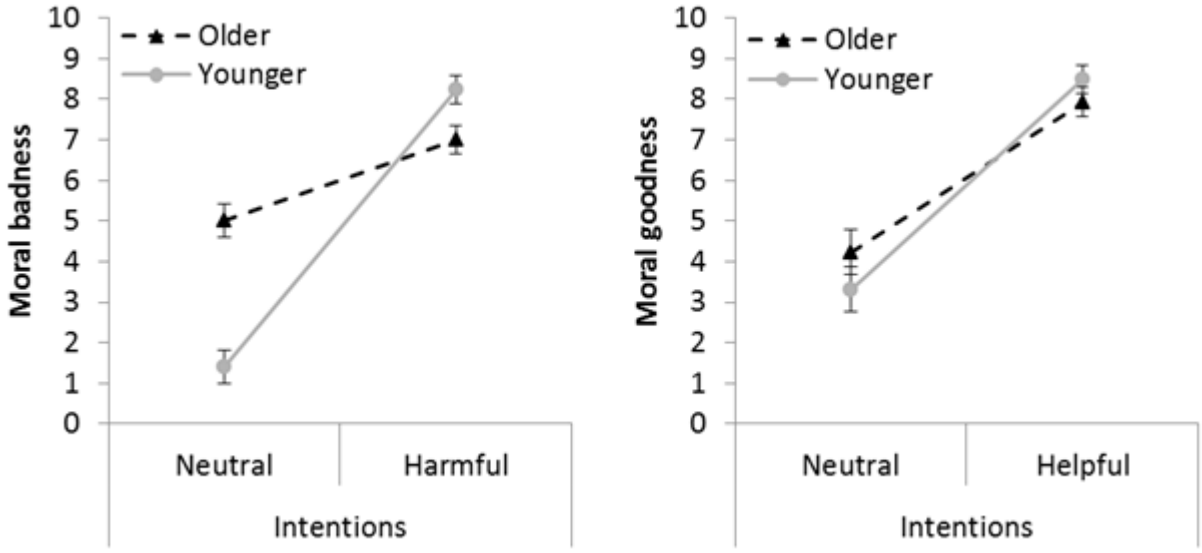

Figure 2. Moral judgment ratings by age (old vs. young) and intentions (neutral vs. valenced). Figure 2a shows moral badness ratings, and Figure 2b moral goodness ratings. Older participants' moral badness ratings were less affected by intentions, there was no age effect for moral goodness ratings. Error bars indicate standard error of the mean.

The analyses also revealed a main effect of outcome, $F(1, 54) = 58.08$, $p < .001$, $f = 1.03$; scenarios involving valenced outcomes received more extreme ratings than ones

involving neutral outcomes. As was the case with intention, and in line with our prediction, this effect was qualified by a significant Age × Outcome interaction, $F(1, 54) = 10.49$, $p = .002$, $f = .44$. Older participants were more influenced by whether an outcome was neutral or valenced than younger participants (see Figure 3). However, unlike the results with intentions, the Age × Outcome interaction was not qualified by a Age × Outcome × Context interaction, $F(1, 54) = 23.19$, $p = .083$, $f = .24$.

Furthermore, we found an Intention × Outcome interaction, $F(1, 54) = 11.05$, $p = .002$, $f = .45$. Intentions exerted a stronger influence for actions that resulted in neutral outcomes, than for actions that resulted in (similarly) valenced outcomes. This is to be expected because intent-based moral evaluations require higher adjustment when the intention conflicts with the outcome (one is neutral while the other is valenced), than when it doesn't (both are neutral, or both are similarly valenced). Finally, the analysis revealed a main effect of context, $F(1, 54) = 4.80$, $p = .033$, $f = .30$. Helpful actions induced more extreme moral judgments (higher ratings) than harmful actions. No other effects were significant.

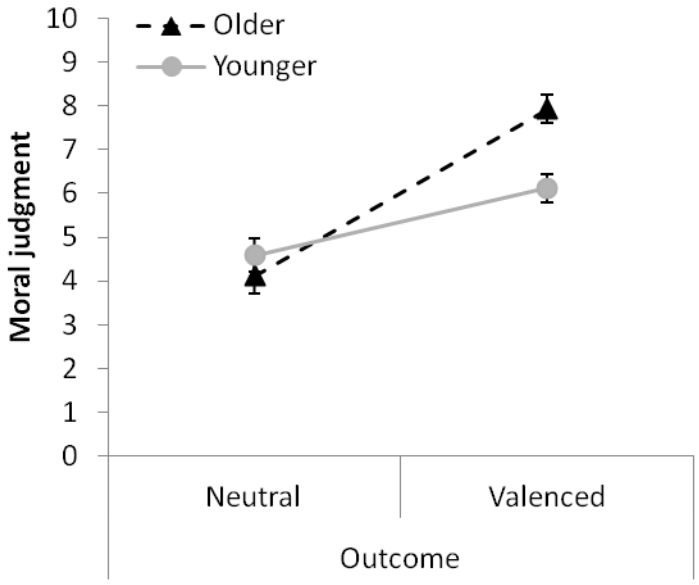

Figure 3. Moral judgment ratings by age (older vs. younger) and outcomes (neutral vs. valenced). Older versus younger participants were more affected by outcome status. Error bars indicate standard error of the mean.

In sum, as expected, older participants weighed less intentions (but mainly for harmful scenarios) and more outcomes than younger participants.

Correlations Between Age, Moral Judgment, Theory of Mind, Empathy, Cognitive Ability, and Executive Function.

Table 2 shows correlations between age, moral judgment, theory of mind, empathy, cognitive ability, and executive function. Age was used as a continuous variable. Age was negatively correlated with moral judgment, theory of mind, cognitive ability, and executive function. Specifically, the higher the age, the less extreme the moral judgments, and the lower the performance in the other tasks. However, age was positively correlated with empathy. That is, older adults scored higher on empathy than younger adults. Moreover, performance in theory of mind, cognitive ability, and executive functions tasks were positively related with one another.

Table 2

Correlations Between Age, Moral Badness, Theory of Mind, Empathy, Cognitive Ability, and Executive Function.

	1	**2**	**3**	**4**	**5**	**6**
1. Age	--					
2. Moral judgment	**–.49****	--				
3. Theory of Mind	**–.54****	**.44****	--			
4. Empathy	**.35***	–.21	–.13	--		
5. Cognitive ability	**–.90****	**.52****	**.52****	**–.32***	--	
6. Executive function	**–.60*****	**.32***	**.32***	–.23	**.60*****	--

Note. Age is a continuous variable. * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

Relationship Between Age, Moral Judgment, Theory of Mind, Empathy, Cognitive Ability, and Executive Function.

We next examined whether age differences in theory of mind, empathy, cognitive ability, and executive function, statistically contribute to age differences in moral judgment. We used 5,000 bootstrapping resamples (Preacher & Hayes, 2008). We created a score to represent the import that intention information has on moral judgments. Specifically, we subtracted the mean moral judgment score assigned to scenarios with neutral intentions from that of scenarios with valenced intentions. Higher scores indicate more intent-based moral judgment. This score was used as the outcome variable in the subsequent analysis. Age was entered as a binary variable (0 = Younger participants, 1 = Older participants). The relationship between age and moral judgment, $b = –3.13$, $t(48) = –4.12$, $p < .001$, 95% CI [–4.650, –1.600,], was reduced after taking into account theory of mind, empathy, cognitive ability, and executive functions, $b = –1.17$, $t(48) = –1.01$, $p = .317$, 95% CI [–3.512, 1.162]. However, only the decline in theory of mind ability significantly reduced the relationship between age and moral judgment, 95% CI [–1.607, –0.073].

Discussion

The present study found that older adults' moral judgment relies less on intentions and more on outcomes than younger adults' moral judgment. The age effect on intentions was more pronounced for harmful than helpful actions, while the age effect on outcomes was unaffected by type of action. Furthermore, in line with previous research, older adulthood was associated with an increase in empathic concern, but a decline in theory of mind, cognitive, and executive function abilities. Importantly, from all these factors, the effect of age on moral judgment was reduced only once we controlled for theory of mind abilities.

This finding lends empirical support to Moran et al.'s (2012) claim that aging differences in this type of moral evaluations are linked to theory of mind.

Why was the age effect on intent-based moral evaluations more pronounced for harmful than for helpful actions? Both neuroimaging and behavioral research suggest that young adults rely more on mental state information when they evaluate harmful rather than helpful actions (e.g., Knobe, 2005; Pizarro, Uhlman, & Salovery, 2003; Young, Scholz, & Saxe, 2011; Young, Nichols, & Saxe, 2010). Similarly, it has been found that young adults weigh more intentions when assigning blame than when assigning praise (Pizzaro et al., 2003). In addition, converging evidence comes from the so-called side-effect effect: Actions with unintended negative side-effects are judged as more intentional than actions with unintended positive side-effects (Knobe, 2005). In line with these results, the young adults in our study placed relatively more weight on intentions when evaluating harmful rather than helpful actions (compare the steepness of the lines representing young adults' moral judgments in Figures 2a and 2b). Returning to the opening question, the reason the aging effect was absent with helpful actions may be because, for such actions, even young adults do not weigh much intentions in their evaluations.

The present findings suggest that older adults' lower reliance on intentions is related to a diminishment in their theory of mind skills. This finding is in line with a number of studies showing that normal aging is associated with theory of mind impairments (for a meta-analysis, see Henry et al., 2013). Although we found no direct evidence for the claim that executive control contributes to the age-related effect on moral judgment, this is likely to be the case. Indeed, many theorists consider executive control to be a critical component of theory of mind skills (e.g., Buon, Seara-Cardoso, Viding, 2016; Henry et al., 2013; Leslie, Friedman & German, 2004). Future studies could investigate further this link by using a

different task to measure executive control and/or by introducing tasks that are known to affect executive functions and examine how these influence moral judgments.

The present results carry implications for everyday judgment and decision making. As stated in the introduction, many important decisions concerning moral issues are made by older adults from the role of judges, politicians, CEOs, and doctors. Consider, for example, an older adult who serves in a jury and should follow the principle of "innocent until proven guilty." A critical component for assigning criminal liability is *mens rea*, that is, that the accused acted with a guilty mind. The present results suggest that older adults may be less concerned with the intentions of the accused and more with the outcomes of their actions, which in the context of criminal trials are dire. Put simply, older adults may find difficult to apply *mens rea* and thus be more likely to convict. This is precisely what a recent study found using data from more than 700 felony trials in Florida (Anwar, Bayer, & Hjalmarsson, 2014). This has tremendous implications for many judicial systems. For example, England and Wales allow people up to the age of 70 to sit in a jury and this limit is set to raise to 75 years. In the US, Federal courts in more than half of the states disallow age-exemptions from jury service.

Limitations

The present study, like all studies, has several limitations. One limitation concerns the materials we used. Research suggests that age-related cognitive declines in social tasks can be improved by using more naturalistic materials (see Light, 1991), such as movies, or materials whose content captures the interest of older adults. Thus, future studies could use movies to convey moral scenarios and/or materials which older adults find interesting. The prediction is that the aging gap may reduce with such materials. However, note that in the current study we purposefully chose scenarios that are likely to capture the interest of our older participants (see for example the harm scenario in Table 1).

A second limitation of the present study relates to its cross-sectional design. The observed differences in moral evaluations may be driven not by age per se but by some other factor that is related to age. For example, they may reflect a cohort effect. It could be that our older participants belonged to a more utilitarian, outcome-focused generation than our younger participants. This would explain why older adults focused relatively more on outcomes and less on intentions. However, in a recent study examining aging effects with sacrificial and non-sacrificial dilemmas, McNair and colleagues (McNair, Okan, Hadjichristidis, & Bruine de Bruin, 2016; see also Arutyunova, Alexandrov, & Hauser, 2016) found that older adults are less utilitarian than younger adults. For example, in relation to the famous footbridge dilemma (Thomson, 1985), older adults were less willing to shove the person off the footbridge to save five other individuals.

Conclusion

The present study shows that older versus younger adults weigh more outcomes and less intentions in their moral evaluations. Importantly, this age-related difference was associated with theory of mind skills: once we controlled for such skills, the association between age and moral evaluation diminished. In a seminal paper, Henrich et al. (2010) noted that most theories in psychology are founded on studies using WEIRD participants, which stands for Western, Educated people from Industrialized, Rich, and Democratic countries. The results from such WEIRD samples, they argued, may not generalize to the world population. The present findings highlight another peculiarity of WEIRD people: they are young adults. Of course, this is nothing new to developmental researchers. Yet, developmental research has predominantly focused on young participants. Perhaps, ironically, aging research on judgment and decision making is still in its infancy. The present findings can inform psychological theories of moral judgment but also public policy for issues such as jury selection.

Footnotes

[1] The sample size was determined by conducting an a-priori power analysis using *G*Power* (Faul, Erdfelder, Lang, & Buchner, 2007) for a repeated-measure analysis of variance including within-between interaction. We used the following estimates: effect size $f = .25$ (medium effect, based on Moran et al. 2013), $\alpha = .05$, power = .95, number of groups = 2, number of measurements = 4, $r = .30$ (estimated), nonsphericity correction $e = 1$. This analysis revealed a minimum sample size of $N = 50$ participants. We recruited more participants as the a-priori power analysis indicated as a precaution of possible drop outs. No interim or stopping rules were applied.

Supplemental Material of Chapter 7

Punishment and Reward Judgments

We analyzed the judgments of punishment and reward. We analyzed these judgments with a 2 (Age: Old vs. Young) × 2 (Intention: Neutral vs. Valenced) × 2 (Outcome: Neutral vs. Valenced) × 2 (Context: Harmful vs. Helpful) mixed-factor ANOVA. Due to space restrictions, we present the full set of results in the supplementary materials. Here we focus on the two main predictions: an interaction between age and intention, and an interaction between age and outcome. The analysis revealed a significant Age × Intention interaction, $F(1, 54) = 18.56$, $p < .001$, $f = .66$. As expected, older as opposed to younger participants were less influenced by whether an intention was neutral or valenced. The analysis also revealed a significant Age × Outcome interaction, $F(1, 54) = 9.70$, $p = .003$, $f = .42$. As expected, older as opposed to younger participants were more influenced by whether an outcome was neutral or valenced. Neither of these effects was further qualified by context, that is, we did not find a significant Age × Intention × Context or Age × Outcome × Context interactions.

Relationship Between Age, Punishment/Reward Judgment, Theory of Mind, Empathy, Cognitive Ability, and Executive Function

We then examined whether age differences in theory of mind, empathy, cognitive ability, and executive function, statistically contribute to age differences in punishment/reward judgment. We used 5,000 bootstrapping resamples (Preacher & Hayes, 2008). We created a total punishment/reward score in the same way as the total moral score described above. Age was entered as a binary variable (0 = Younger participants, 1 = Older participants). The relationship between age and punishment/reward judgment, $b = –3.19$, $t(48) = –4.17$, $p < .001$, 95% CI [–4.738, –1.650], was reduced after taking into account

theory of mind, empathy, cognitive ability, and executive functions, $b = –0.48$, $t(48) = –0.41$, $p = .685$, 95% CI [–2.834, 1.880]. However, only the decline in cognitive ability significantly reduced the relationship between age and punishment/reward judgment, 95% CI [–4.122, –0.278].

GENERAL DISCUSSION AND PERSPECTIVES

General Discussion and Perspectives

1. Main Findings

The present work focused on infants' expectations of obedience and the development of a mature moral judgment mainly based on the agent's intention assessment. As argued in the general introduction, both the respect for authority or leadership and the evaluation of the intention beyond the action are crucial aspects of our morality. The conformity to a certain set of rules and the obedience to the authority reside at the very core of human morality; as the child grows, he or she begins to fully appreciate the complexity of his or her fellows' morality. The child evaluates not only whether others actually respected the rules, but also whether they intended to respect the rules. Compared to action consequences, intention is arguably a much more informative cue to predict the outcome of future interactions and to maintain the cooperation within a group.

1.1 Expectations of Obedience

In Part 1 I presented a research that addressed the question whether in the second year of life infants already possess the ability to represent the leader-followers relationship and whether they expect a leader's instruction but not a bully's instruction to be obeyed by a group of subordinates. For the first time, we reported evidence that 21-month-olds are able to discriminate between two distinct types of dominance, leadership or authority and physical dominance or bullying. This distinction is crucial for the understanding of human morality, and it may represent a basis for the representation of our complex social words that mainly consists in a set of hierarchical social structures evolved in order to establish and maintain order, mutual respect and cooperation between individuals.

While a leader can be defined as an individual whose source of power is deemed rightful or it is spontaneously accepted by the subordinates, a bully, at least in the present context, is an individual that tends to prevail in conflicting situations, primarily by means of physical coercion or intimidation. Following these definitions, we see that leaders could have a moral authority, but bullies just exercise a coercive power when their forces or skills permit them to overhang subordinates. Our data suggested that infants seem to understand these complex dynamics. They expected subordinates to obey to an absent leader, but not to an absent bully—the bully's influence is indeed constrained by the fact that he used physical force to gain the subordinates' initial compliance, and infants subsequently expected that subordinates would not be influenced by the bully's command when he is not present to control. With a further experiment, we also excluded that infants' expectations of obedience in leader condition was due to a general positive interaction between characters and not to leadership. Our findings add an important building block to the understanding of the early mechanisms for representing socio-moral situations and, as I will argue in the perspective session, they may be considered an initial piece of evidence for an early-developing 'naïve politics'.

1.2 Intent-based Moral Reasoning in Children

Part 2 is focused on the children's developing intent-based moral judgment. In particular, in Chapter 2, we aimed to investigate at what age children produce an intent-based goodness judgment of agents that either attempted but failed or accidentally helped someone. We reported that, when presented with a verbal moral task asking to evaluate the goodness or badness of some agents, children aged 5-6, but not 4 year-olds, attend to agents' intentions more than to actions outcomes. This developmental shift, also known as 'the outcome-to-

intent shift', is a major one in the moral development (e.g., Cushman et al., 2013; Killen et al., 2011; Margoni & Surian, 2017; Nobes et al., 2009; Piaget, 1932). The child acquires a full-blown capacity to express a moral judgment that is integrated with the understanding of others' mental state. A further and related question we addressed was whether this shift reflects a conceptual change (in the concept of moral goodness) or ancillary changes occurring outside the moral domain, for instance in theory of mind abilities or executive functioning. Our results do not support the hypothesis of a conceptual change in the moral domain, and overall are more consistent with the hypothesis that the shift reflects changes occurring outside the moral domain.

Although future work is needed in order to conclude whether the outcome-to-intent shift reflects a conceptual change or not, in Chapter 4 we further argued in favor of a continuity hypothesis, that is, a conceptual change is both unlikely and unnecessary to explain current findings. Indeed, we know from infant cognition literature that already at the end of the first year of life the child shows to be able to attend to intention in his or her socio-moral expectations and preferences (Dunfield & Kuhlmeier, 2010; Hamlin, 2013; Lee et al., 2015). Infant development literature uses mainly spontaneous-response tasks or simple elicited-response tasks that often decrease the processing demand compared to verbal tasks used to investigate moral judgment in preschoolers (Baillargeon et al., 2015). We then argued that 4-year-olds show an outcome-based moral judgment not because they do not yet have developed an intent-based concept of moral goodness or badness, but because the task with which they are typically presented is too demanding in terms of cognitive processing. In order to produce an intent-based judgment in a verbal elicited-response task, the child needs to suppress the information concerning the action outcomes and subsequently select an intent-based response. This cognitive processing requires executive functioning skills that at 4 years may be not yet fully developed. Therefore, what really develops is the child' executive

functioning ability, and this development—given the nature of the tasks being used in the children literature—likely produces the outcome-to-intent shift.

In Chapter 5, then, we reviewed recent works linking mental state understanding and moral judgment in individuals with autistic spectrum disorders. We concluded that this clinical population encounters some difficulties in developing a full-blown intent-based judgment, likely because of the impairment in the mental state understanding. The research on autism, then, proves useful to determine which component is required for the development of a mature intent-based moral judgment. Executive functioning skills (Chapter 4) and mental state reasoning (Chapter 5) are crucial components that are required to elaborate a moral judgment, and could explain a large part of our moral development. A further evidence for this line of reasoning was provided in Chapter 7, Part 3.

Finally, in the third chapter, I presented a study that we conducted in order to expand on the issue of the preschoolers' development of intent-based moral judgment. We asked whether the moral evaluations and preferences of children aged 4 to 5 already rely on two distinct types of intentions. On the one hand, a biocentric intention is defined as the intention to preserve nature because of the nature intrinsic value; on the other hand, an anthropocentric intention is the intention to preserve nature because it helps humans' interests (Kahn & Friedman, 1995; White, 1967). Here we investigated not only whether 5-year-olds rely on intention when judging the morality of some actions or agents, but we investigated whether they are able to show a preference between intentions that reflect two contrasting moral views or ways of extending ethics to natural entities.

1.3 Intent-based Moral Reasoning in Adults

In Chapter 6 and 7, I reported two studies that investigated the intent-based moral judgment respectively in adults and older adults. Current models of adults' moral judgment do not address or remain unclear on three main issues: first, we still do not have convincing evidence for concluding that the cognitive processing mechanisms underlying the attribution of moral goodness are the same of those underlying the well-studied attribution of moral badness or wrongness; second, current models diverge on the role and weight of negligence information; third, little attention has been paid on how older people' moral judgment works.

In Chapter 6, we showed that in adults the attribution of both moral rightness and wrongness are based mainly on an assessment of the agent's intention. Moreover, we clarified that also negligence plays a significant role, albeit a marginal one if compared to the role played by intention. Consequences, instead, affected more deserved reward and punishment judgments. We therefore suggest that a current influential dual-process model of moral judgment (Cushman, 2008) should be generalized to account also for rightness judgment, but needs a minor revision to account for small differences we detected between rightness and wrongness judgments, and for the marginal but still significant role played by negligence information in shaping our moral evaluations.

A further integration is needed for the existing modeling of the development of intent-based moral judgment. In Chapter 7, we presented a study revealing an 'intent-to-outcome shift' in old age. We reported that older adults' moral judgments rely more on action outcomes and less on agent's intentions compared to the judgments of younger adults. This shift in later years was found to be more pronounced in the evaluation of harming actions than in the evaluation of helping actions. Indeed, we found that when judging the moral badness, older adults' judgments undergo a full-blown shift from intention to outcome, while

older adults' moral evaluations of goodness were only more outcome-based and not less intent-based compared to younger adults' evaluations.

Moreover, we reported that the age-related change in moral judgment was associated with the well-known decrease in theory of mind abilities occurring in 'golden years'. These results can also be relevant for the issue we addressed in Part 2 whether the developmental shift occurring in moral judgment during childhood reflects ancillary changes occurring outside the moral domain or a conceptual change within the moral domain. Our evidence on older adults' moral judgment highlights the importance of ancillary changes in theory of mind (and executive functioning skills, indeed a necessary component for theory of mind) in shaping the development of intent-based moral judgment throughout our life. A description of the development of moral judgment consistent with our data would insist on the changes occurring in the executive functioning. These changes may account for the outcome-to-intent shift during childhood and for the intent-to-outcome shift during later years.

2. Perspectives

With respect to the research on infants' understanding of dominance relationships, I predict that in the next years the field will move forward at high speed. A few studies have been conducted so far, and a lot of exciting questions remained to be investigated. Previous research showed that infants are able to use the physical size of characters and the numerical size of groups to predict the outcome of a conflict between two characters (Thomsen et al., 2011; Pun et al., 2016). However, other cues may be relevant to predict the outcome of a conflict. A first one is hunger. Typically, hungry people value the resources more than satiated individuals, especially when the resources are food and they are scarce. Evidence

from animal research have also shown that individuals in need of a particular resource spontaneously emerges as leaders and often coordinate foraging (Krause, 1993; Rands, Cowlishaw, Pettifor, Rowcliffe, & Johnstone, 2003). Do infants expect hungry individuals to prevail over non-hungry individuals? Do they expect hungry individuals to search longer for resources than non-hungry individuals?

Mascaro & Csibra (2012) defined 'social dominance' as the tendency to prevail in competitive situations. However, finding that a hungry subordinate would actually prevail over a dominant because of his or her stronger motivation to fight could constrain the definition initially proposed by Mascaro & Csibra (2012). A second cue may be strength. Physical size could be less accurate than strength for predicting which individual will prevail. Do infants expect individuals with greater strength to prevail? And, do infants rely more on size or more on strength to predict the outcome of a conflict?

Moreover, the line of research regarding infants' expectations of obedience, that we started with the study presented in Chapter 1, could be followed by two main projects. A first project could be devoted to the understanding of which cues determine or form the representation of leadership during early infancy. Of course, leadership is a social relation, so we may want, more specifically, to think about the leader-followers representation. Leadership is also a complex phenomenon, and I believe that no single cue can determine such a representation. For this reason, a set of cues is likely to be the cause of the leader-followers mental representation.

As we argued, leaders serve the function of maintaining the cooperation and cohesion in the group; moreover, they facilitate the cultural evolution of the group through a mechanism of imitation (e.g., Henrich & Gil-White, 2001; Van Vugt, 2006). Then, a first element that may determine the mental representation of leadership could be the fact that subordinates

show respect towards the leader and the leader respects and cares for subordinates. So, a mutual respect between parties could be an important element of this particular social relationship. A second element could be that subordinates tend to imitate the leader. A third set of elements could be related to the material features that help identifying a chief, or someone that is in charge, such as a higher spatial position and a larger space. Overall, future research should identify which of these or other characteristics give rise to the infants' expectations of obedience towards an absent dominant.

A second project could be proposed in order to investigate whether a leader figure, as opposed to a bully figure, has a positive impact on infants' learning abilities. As we have seen, some evolutionary psychologists posited that subordinates might be motivated to follow leaders and to allow them some privileges in exchange of the possibility to learn from them (e.g., Aidar, 1989; Berger et al., 1972; Henrich & Gil-White, 2001). Indeed, leaders are often skilled individuals, and previous research already showed that infants' imitation is influenced by the reliability of the model (Zmyj, Buttelmann, Carpenter, & Daum, 2010). According to this line of reasoning, we can hypothesize that leaders attract more attention than bullies. Therefore, children's ability to reproduce a given sequence of instructions (e.g., how to build a toy) could be improved if the instructor has the qualities of a leader vs. a bully.

Finally, the field could be really moved forward by studying what I call here the "naïve politics" in early infancy. Seminal work in this eventual new area of research may be done by simplifying and opposing two main well-known views of politics (and economy): communism and liberalism. A first step would be to investigate infants' representation of (private) ownership. Do infants expect a character that was holding a toy to prevail to a character that did not hold the toy when they conflict over it? Do infants expect owners to search longer for their objects than non-owners? And what about their implicit evaluations? Do infants have preferences that suggest the normative idea that property should be private?

Recent works have shown that infants are able to understand transfer-based interactions between agents, both taking and giving actions (e.g., Geraci & Surian, 2011; Meristo & Surian, 2013; Schöppner, Sodian, & Pauen, 2006; Tatone et al., 2015). However, so far, no direct evidence of an early-emerging concept of private property has been reported.

A second step would be to investigate whether infants prefer fairness over individual freedom, or vice versa. Finally, a third step would be to simulate, for instance with a puppet show, the dynamics of the workers exploitation. Hypothesizing that infants would watch and understand a character playing the role of a master that do not redistribute equitably among the workers or slaves the resources derived from their work, would infants also prefer the worker over the master? And, would infants expect an equal or unequal distribution of the resources derived from the workers' job? Indeed, it has been shown that children at around the age of three are extremely willing to share resources equitably in collaborative activities, and in those situations they are more willing to share than in situations in which simply there is an abundance of resources or the child worked in parallel with others (Hamann, Warneken, Greenberg, & Tomasello, 2011). This important piece of evidence suggests that the child possesses a complex sense of justice, which is constrained by some relevant aspects of the situation, such as the collaborative nature of the job or activity that has been done. However, future investigation should reveal whether this sense of justice with respect to collaborative activities is innate or culturally learned.

Coming now to the research on the development of intent-based moral judgment, two main future perspectives are considered here. First, more research is needed to address the issue whether developmental changes in moral judgment reflect conceptual change or ancillary changes occurring outside the moral domain, such as in theory of mind or executive functioning skills. Although in the present dissertation I offered some arguments in favor of conceptual continuity during development, current data do not allow any strong conclusion

on this important issue. Clever studies that would contribute to the investigation of the role of executive function and theory of mind in the development of moral judgment should then be in order.

Second, future studies should explore the possibility to integrate existing processing models of moral judgment by adding to them also a process that analyzes whether a given action, or intention, respects the authority's mandate. Indeed, current models neglect the role that obedience to authority and conformity could and likely play in our everyday moral evaluations. Both Piaget (1932) and the social domain theorists (e.g., Killen & Smetana, 2015; Smetana, 2006; Turiel, 1983) carefully investigated the authority-dependency of moral rules in the child's judgments. However, so far, scarce work has been done in order to study the information processing aspects underlying people's moral evaluations of moral actions and authority's mandates.

3. Conclusion

The present dissertation collects several works overall aimed at investigating some crucial aspects of the development of moral evaluation. I presented experiments on both the infants' capacity to distinguish between leaders and bullies, and the development of the mature processing of others' intentions during the production of a moral judgment. Both aspects are crucial for morality, that is, the respect for the rules. On the one hand, respect for authority constitutes a first pillar of morality. We showed that 21-month-olds already hold the expectation that subordinates obey to leaders, who have the moral authority, but not to bullies, who simply use physical force or coercion to dominate and thus are devoid of any moral authority. On the other hand, a more complex understanding of others leads children to

evaluate not only whether the action or its outcomes are violating some rules, but also whether the agents' intentions are consistent with the authority's mandates and the moral rules. We showed that an outcome-to-intent shift in moral reasoning occurs during preschool years also when children are presented with helping actions and are asked to evaluate the moral goodness of the character. Moreover, we found that older adults return to attend to outcome instead of intention when judging the morality of an agent. We argued that both shifts likely reflect general changes occurring outside the moral domain.

References

Abelard, P. (1971; originally written in the XII century). *Ethics.* Oxford: Oxford University Press.

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development, 15,* 1-16.

Aidar, J. (1989). *Great leaders.* Guilford, UK: Talbot Press.

Anderson, C., & Kilduff, G. J. (2009). The pursuit of status in social groups. *Current Direction in Psychological Science, 18,* 295-298.

Andrighetto, L., Baldissarri, L., Lattanzio, S., Lattanzio, S., Loughnan, S., & Volpato, C. (2014). Human-itarian aid? Two forms of dehumanization and willingness to help after natural disasters. *British Journal of Social Psychology, 53,* 573-584.

Anwar, S., Bayerand, P., & Hjalmarsson, R. (2014). The role of age in jury selection and trial outcome. *Journal of Law and Economics, 57,* 1001–30.

Amieva, H., Phillips, L., & Della Sala, S. (2003). Behavioral dysexecutive symptoms in normal aging. *Brain and Cognition, 53,* 129–32.

Armon, C., & Dawson, T. (1997). Developmental trajectories in moral reasoning across the life span. *Journal of Moral Education, 26,* 433-453.

Armsby, R. E. (1971). A reexamination of the development of moral judgments in children. *Child Development, 42,* 1241-1248.

Arsenio, W. (1988). Children's conceptions of the situational affective consequences of sociomoral events. *Child Development, 59,* 1611-1622.

Arutyunova, K. R., Alexandrov, Y. I., & Hauser, M. D. (2016). Sociocultural influences on moral judgments: East–West, male–female, and young–old. *Frontiers in Psychology, 7,* 1334.

Aslin, R. N. (2000). Why take the cog out of infant cognition? *Infancy, 1,* 463-470.

Astuti, R., & Bloch, M. (2015). The causal cognition of wrong doing: Incest, intentionality, and morality. *Frontiers in Psychology, 6,* 136.

Aureli, F., & de Waal, F. (2000). *Natural conflict resolution.* Berkeley: University of California Press.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211,* 1390-1396.

Bailey, P. E., & Henry, J. D. (2008). Growing less empathic with age: Disinhibition of the self-perspective. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 63,* 219–226.

Baillargeon, R., Scott, R., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14,* 110-118.

Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K., Wu, D., & Bian, L. (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer, P. R. Shaver (Eds.), E. Borgida, & J. A. Bargh (Assoc. Eds.), *APA handbook of personality and social psychology: Vol. 1. Attitudes and social cognition* (pp. 79-150). Washington, DC: American Psychological Association.

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.

Banaji, M. R., & Gelman, S. A. (2013). *Navigating the social world.* New York: Oxford University Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition, 21,* 37-46.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry, 38,* 813–822.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (2000). *Understanding other minds: Perspectives from developmental cognitive neuroscience*. Oxford: Oxford University Press.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42,* 241–251.

Barrett, H. C., Bolyanatz, A., Crittenden, A., Fessler, D., et al. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Pnas, Early Edition*.

Bass, B. M. (1990). *Bass and Stogdill's handbook of leadership: Theory, research, and managerial applications*. New York: Free Press.

Baumard, N., André, J., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36,* 59-122.

Begeer, S., Gevers, C., Clifford, P., Verhoeve, M., Kat, K., Hoddenbach, E., & Boer, F. (2011). Theory of mind training in children with Autism: A randomized controlled trial. *Journal of Autism and Developmental Disorders, 41*, 997-1006.

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

Benes, F. (2001). The development of frontal cortex: The maturation of neurotransmitter systems and their interactions. In C. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 79-92). Cambridge, MA: MIT Press.

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

Berger, J., Cohen, B. P., & Zelditch, M. (1972). Status characteristics and social interaction. *American Sociological Review, 37,* 241-255.

Berger, J., Rosenholtz, S. J., & Zelditch, M. (1980). Status organizing processes. *Annual Review of Sociology, 6,* 479-508.

Blair, J. (1996). Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders, 26,* 571-579.

Blair, J. (1999). Psychophysiological responsiveness to the distress of others in children with autism. *Personality and Individual Differences, 26,* 477-485.

Bloom, P. (2012). Religion, morality, evolution. *Annual Review of Psychology, 63,* 179-199.

Bloom, P. (2013). *Just babies: The origins of good and evil.* New York: Crown Publisher.

Bloom, P., & Wynn, K. (2016). What develops in moral development? In D. Barner & A. S. Baron (Eds.), *Core knowledge and conceptual change* (pp. 347-364). New York: Oxford University Press.

Boehm, C. (1999). *Hierarchy in the forest.* London: Harvard University Press.

Bohner, G., Bless, H., Schwarz, N., & Strack, F. (1988). What triggers causal attribution? The impact of valence and subjective probability. *European Journal of Social Psychology, 18,* 335-345.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Bostyn, D. H., & Roets, A. (2016). The morality of action: The asymmetry between judgments of praise and blame in the action-omission effect. *Journal of Experimental Social Psychology, 63,* 19-25.

Bowler, D. M. (1992). “Theory of mind” in Asperger’s syndrome. *Journal of Child Psychology and Psychiatry, 33,* 877-893.

Brandone, A., & Wellman, H. M. (2009). You can’t always get what you want: Infants understand failed goal-directed actions. *Psychological Science, 20,* 85–91.

Bronson, W. C. (1975). Developments in behavior with age mates during the second year of life. In M. Lewis & L. A. Rosenblum (Eds.), *Friendship and peer relations* (pp. 131-152). New York: Wiley.

Brown, D. (1991). *Human universals*. Boston: McGraw-Hill.

Brownell, C. A., Iesue, S. S., Nichols, S. R., & Svetlova, M. (2013). Mine or yours? Development of sharing in toddlers in relation to ownership understanding. *Child Development, 84,* 906-920.

Buon, M., Dupoux, E., Jacob, P., Chaste, P., Leboyer, M., & Zalla, T. (2013). The role of causal and intentional judgments in moral reasoning in individuals with high functioning autism. *Journal of Autism and Developmental Disorders, 43,* 458-470.

Buon, M., Seara-Cardoso, A., & Viding, E. (2016). Why (and how) should we study the interplay between emotional arousal, Theory of Mind, and inhibitory control to understand moral cognition? *Psychonomic Bulletin & Review,* Advance on-line publication.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112,* 337-342.

Carpenter, M., Pennington, B. F., & Rogers, S. J. (2001). Understanding of others’ intentions in children with autism. *Journal of Autism and Developmental Disorders, 31,* 589-599.

Caplan, M., Vespo, J., Pedersen, J., & Hay, D. F. (1991). Conflict and its resolution in small groups of one- and two-year-olds. *Child Development, 62,* 1513-1524.

Casey, P. J., & Scott, K. (2006). Environmental concern and behavior in an Australian sample within an ecocentric-anthropocentric framework. *Australian Journal of Psychology, 58,* 57-67.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125,* 1839-1849.

Chakroff, A., & Young, L. (2015). How the mind matters for morality. *AJOB Neuroscience, 6,* 41-46.

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

Channon, S., Fitzpatrick, S., Drury, H., Taylor, I., & Lagnado, D. (2010). Punishment and sympathy judgments: Is the quality of mercy strained in Asperger's Syndrome? *Journal of Autism and Developmental Disorders, 40,* 1219-1226.

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology, 104,* 103-125.

Choe, S. Y., & Min, K.-H. (2011). Who makes utilitarian judgments? The influences of emotions on utilitarian judgments. *Judgment and Decision Making, 6,* 580–592.

Choi, Y., & Luo, Y. (2015). 13-month-olds' understanding of social interactions. *Psychological Science, 26,* 274-283.

Clark-Barrett, H., Bolyanatz, A., Crittenden, A., Fessler, D., et al. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences, 113,* 4688-4693.

Clayton, S. (1998). Preferences for macrojustice versus microjustice in environmental decisions. *Environmental and Behavior, 30,* 162-183.

Coie, J. D., & Dodge, K. A. (1983). Continuities and changes in children's social status: A five-year longitudinal study. *Merrill-Palmer Quarterly, 29,* 261-282.

Coleman, J. S., & Temple, S. A. (1996). On the prowl. *Wisconsin Natural Resources Magazine, 20,* 4-8.

Corraliza, J. A., Collado S., & Bethelmy, L. (2013). Spanish version of the New Ecological Paradigm Scale for children. *The Spanish Journal of Psychology, 16,* E27 doi:10.1017/sjp.2013.46.

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: New perspective on cognition and motivation. *Annual Review of Psychology, 64,* 201-229.

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

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, 107,* 17433–17438.

Cummins, D. (2006). Dominance, status, and social hierarchies. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 676-697). Hoboken, NJ: Wiley.

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

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17,* 273-292.

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology, 6,* 97-103.

Cushman, F. A. (unpublished). The role of mental state attribution in moral judgment: A dissociation between cases of help and harm.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*, 6-21.

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

Dahl, R. A. (1957). The concept of power. *Behavioral Science, 2,* 201-215.

D'Arcy, E. (1963). *Human acts: An essay in their moral evaluation.* Oxford: Clarendon.

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

Darwin, C. (1859/1982). *The origin of species.* Berkeley: Penguin Classics.

Davis, M., H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10,* 85.

Decety, J., & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in Cognitive Sciences, 18,* 337–339.

Descartes, R. (1637/1970). *Discourse on method.* Trans. E. S. Haldane & G. R. Ross, Cambridge: Cambridge University Press.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64,* 135-168.

Diamond, J. (1998). *Guns, germs and steel.* London: Vintage.

Doris, J. M. (Ed.). (2010). *The moral psychology handbook.* Oxford: Oxford University Press.

Dunfield, K. A., & Kuhlmeier, V. A. (2010). Intention-mediated selective helping in infancy. *Psychological Science, 21,* 523-527.

Eckerman, C. O., Davis, C. C., & Didow, S. M. (1989). Toddlers' emerging ways of achieving social coordination with a peer. *Child Development, 60,* 440-453.

Enzle, M. E., & Hawkins, W. L. (1992). A priori actor negligence mediates a posteriori outcome effects on moral judgment. *Journal of Experimental Social Psychology, 28,* 169-185.

Fadda, R., Parisi, M., Ferretti, L., Saba, G., Foscoliano, M., Salvago, A., & Doneddu, G. (2016). Exploring the role of Theory of Mind in moral judgment: The case of children with autism spectrum disorder. *Frontiers in Psychology, 7,* 523.

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

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Feltz, A., & Cokely, T. (2013). Virtue or consequences: The folk against pure evaluational internalism. *Philosophical Psychology, 26*, 702-717.

Fincham, F. D., & Roberts, C. (1985). Intervening causation and the mitigation of responsibility for harm doing: The role of limited mental capacities. *Journal of Experimental Social Psychology, 21,* 178-194.

Finkel, N., & Groscup, J. L. (1997). When mistakes happen: Commonsense rules of culpability. *Psychology, Public Policy, and Law, 3,* 65-125.

Fisher, N., & Happe, F. (2005). A training study of Theory of Mind and executive function in children with autistic spectrum disorders. *Journal of Autism and Developmental Disorders, 35*, 757-771.

Fiske, A. P. (1991). *Structures of social life.* New York: Free Press.

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review, 99,* 689-723.

Fiske, A. P., & Haslam, N. (2005). The four basic social bonds: Structures for coordinating interaction. In M. Baldwin (Ed.), *Interpersonal cognition* (pp. 267-298). New York: Guildford Press.

Fiske, A. P., & Rai, T. S. (2014). Violence for goodness' sake. *New Scientist, 224,* 30-31.

Fiske, S. T. (2010). Interpersonal stratification: Status, power, and subordination. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology.* Hoboken, NJ: Wiley.

Folstein, M., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198.

Frith, U., Morton, J., & Leslie, A. (1991). The cognitive basis of a biological disorder: Autism. *Trends in Neuroscience, 14,* 433-438.

Gagnon Thompson, S. C., & Barton, M. (1994). Ecocentric and anthropocentric attitudes toward the environment. *Journal of Environmental Psychology, 14,* 149-157.

Gazes, R. P., Hampton, R. R., & Lourenco, S. F. (2015). Transitive inference of social dominance by human infants. *Developmental Science,* doi: 10.1111/desc.12367.

Gebert, D., Heinitz, K., & Buengeler, C. (2016). Leaders' charismatic leadership and followers' commitment – The moderating dynamics of value erosion at the societal level. *The Leadership Quarterly, 27,* 98-108.

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., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences, 7,* 287-292.

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

Gleichgerrcht, E., Torralva, T., Rattazzi, A., Marenco, V., Roca, M., & Manes, F. (2013). Selective impairment of cognitive empathy for moral judgment in adults with high functioning autism. *SCAN, 8,* 780-788.

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PloS One, 8,* e60418.

Goldstein, J. (2011, January 18). Life tenure for federal judges raises issues of senility, dementia. Retrieved from https://www.propublica.org/article/life-tenure-for-federal-judges-raises-issues-of-senility-dementia

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives use different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029-1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101,* 366-385.

Grant, C., Boucher, J., Riggs, K., & Grayson, A. (2005). Moral understanding in children with autism. *Autism, 9,* 317-331.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23,* 101-124.

Greene, J. D. (2013). *Moral tribes.* New York: Penguin Press.

Greene, J., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105-2017.

Grueneich, R. (1982). Issue in the development study of how children use intention and consequence information to make moral evaluations. *Child Development, 53,* 29-43.

Guinote, A., & Vescio, T. K. (2010). *The social psychology of power*. New York: Guilford Press.

Gülgöz, S., & Gelman, S. A. (in press). Who's the boss? Concepts of social power across development. *Child Development*.

Gvozdic, K., Moutier, S., Dupoux, E., & Buon, M. (2016). It's not only mental states that count! A specific inhibitory reinforcement on children's ability to generate intent-based moral judgment. *Frontiers in Psychology, 7,* 190.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65,* 613-628.

Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature, 476,* 328-331.

Hamilton, A. F. (2009). Goals, intentions and mental states: Challenges for theories of autism. *Journal of Child Psychology and Psychiatry, 50,* 881-892.

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition, 128,* 451-474.

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,* 557-559.

Hamlin, J. K., Wynn, K., & Bloom, P. (2010). 3-month-olds show a negativity bias in their social evaluations. *Developmental Science, 13,* 923-929.

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments with preverbal infants and a computational model. *Developmental Science, 16,* 209–226.

Hand, J. L. (1986). Resolution of social conflicts: Dominance, egalitarianism, spheres of dominance, and game theory. *Quarterly Review of Biology, 61,* 201-220.

Hart, H. L. (1968). *Punishment and responsibility.* Oxford: Clarendon Press.

Haupt, A., & Uske, T. (2012). The asymmetry of praise and blame: Distinguishing between moral evaluation effects and scenario effects. *Journal of Cognition and Culture, 12,* 49-66.

Hawley, P. H. (1999). The ontogenesis of social dominance: A strategy-based evolutionary perspective. *Developmental Review, 19,* 97-132.

Hawley, P. H. (2002). Social dominance and prosocial and coercive strategies of resource control in preschoolers. *International Journal of Behavioral Development, 26,* 167-176.

Hay, D. F. (1984). Social conflict in early childhood. In G. Whitehurst (Ed.), *Annals of child development* (Vol. 1, pp. 1-44). Greenwich, CT: JAI.

Hay, D. F., & Ross, H. S. (1982). The social nature of early conflict. *Child Development, 53,* 105-113.

Heaton, R. K. (1995). *Wisconsin Card Sorting manual.* Odessa, FL: Psychology Assessment Resources.

Hegel, G. W. F. (1807/1977). *Phenomenology of Spirit.* Trans. A. V. Miller, Oxford: Oxford University Press.

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

Heinrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior, 22,* 165-196.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences, 33,* 61–83; discussion 83–135.

Helwig, C., Hildebrandt, C., & Turiel, E. (1995). Children's judgments about psychological harm in social context. *Child Development, 66,* 1680-1693.

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging, 28,* 826–839.

Hepach, R., Haberl, K., Lambert, S., & Tomasello, M. (2016). Toddlers help anonymously. *Infancy,* 10.1111/infa.12143.

Hill, K., & Hill, C. (1977). Children's concept of good and bad behavior. *Psychological Reports, 41,* 955-958.

Hirschfeld, L. A. (1999). Naïve sociology. *The MIT Encyclopedia of the Cognitive Sciences* (pp. 579-580). Cambridge, MA: MIT Press.

Hobbes, T. (1651/1982). *Leviathan.* Penguin.

Hobson, J., Harris, R., Garcia-Péres, R., & Hobson, R. (2009). Anticipatory concern: A study in autism. *Developmental Science, 12,* 249-263.

Hoffman, M. L. (1991). Empathy, social cognition, and moral action. In WM Kurtines e JL Gewirtz (Eds.), *Handbook of moral behavior and development theory* (pp. 275-301). Hillsdale, NJ: Erlbaum Associates.

Hogan, R., Curphy, G. J., & Hogan, J. (1994). What we know about leadership. *American Psychologist, 49,* 493-504.

Holmberg, M. C. (1980). The development of social interchange patterns from 12 to 24 months. *Child Development, 51,* 448-456.

Howe, D., Kahn, P. H., & Friedman, B. (1996). Along the Rio Negro: Brazilian children's environmental views and values. *Developmental Psychology, 32,* 979-987.

Hume, D. (1740/1978). *A treatise on human nature.* London: Clarendon.

Hussar, K. M., & Horvath, J. C. (2011). Do children play fair with mother nature? Understanding children's judgments of environmentally harmful actions. *Journal of Environmental Psychology, 31,* 309-313.

Huttenlocher, P., & Dabholkar, A. (1997). Developmental anatomy of prefrontal cortex. In N. Krasnegor, G. Lyon, & P. Goldman-Rakic (Eds.), *Development of the prefrontal cortex: Evolution, neurobiology, and behavior* (pp. 69-83). Baltimore: Brookes Publishing.

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

Jambon, M., & Smetana, J. D. (2013). Moral complexity in middle childhood: Children's evaluations of necessary harm. *Developmental Psychology, 50,* 22-33.

Joyce, R. (2006). Is human morality innate? In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 2, pp. 257-279). New York: Oxford University Press.

Kahane, G., Everett, J., Earp, B., Farias, M., & Savulescu, J. (2014). Utilitarian judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134,* 193–209.

Kahn, P. H. (1992). Children's obligatory and discretionary moral judgments. *Child Development, 63,* 416-430.

Kahn, P. H. (1997). Children's moral and ecological reasoning about the Prince William Sound oil spill. *Developmental Psychology, 33,* 1091-1096.

Kahn, P. H., & Friedman, B. (1995). Environmental views and values of children in an inner-city Black community. *Child Development, 66,* 1403-1417.

Kahn, P. H., & Lourenco, O. (2002). Water, air, fire, and earth: A developmental study in Portugal of environmental moral reasoning. *Environment and Behavior, 34,* 405-430.

Kaiser, F. G., Ranney, M., Hartig, T., & Bowler, P. A. (1999). Ecological behavior, environmental attitude, and feelings of responsibility for the environment. *European Psychologist, 4*, 59-74.

Kant, I. (1785/1959). *Foundation of the metaphysics of morals.* Indianapolis: Bobbs-Merrill.

Kant, I. (1788/2002). *The critique of practical reason.* Trans. W. S. Pluhar, Cambridge: Hacklett.

Kaplan, H. S., & Gangestad, S. W. (2005). Life history theory and evolutionary psychology. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (pp. 68-95). Wiley.

Karniol, R. (1978). Children's use of intention cues in evaluating behavior. *Psychological Bullettin, 85*, 76-85.

Keasey, C. B. (1978). Young children's attribution of intentionality to themselves and others. *Child Development, 48*, 261–264.

Kellert, S. R. (1985). Attitudes towards animals: Age related development among children. *Journal of Environmental Education, 16*, 29-39.

Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, affect, and the moral/conventional distinction. *Mind & Language, 22*, 117-131.

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

Killen, M., & Smetana, J. (2008). Moral judgment and moral neuroscience: Intersections, definitions, and issues. *Child Development Perspectives, 2*, 1–6.

Killen, M., & Smetana, J. (2015). Origins and development of morality. In M. E. Lamb (Ed.), *Handbook of child psychology and developmental science*, Vol. 3, 7th ed. (pp. 701-749). New York: Wiley-Blackwell.

Killen, M., & Turiel, E. (1991). Conflict resolution in preschool social interactions. *Early Education and Development, 2*, 240-255.

King, A. J., Johnson, D. D., & Van Vugt, M. (2009). The origins and evolution of leadership. *Current Biology, 19,* 911-916.

King, M. (1971). The development of some intention concepts in young children. *Child Development, 42,* 1145-1152.

Knauft, B. M. (1991). Violence and sociality in human evolution. *Current Anthropology, 23,* 391-428.

Knight, R., & Stuss, D. (2002). Prefrontal cortex: The present and the future. In D. Stuss & R. Knight (Eds.), *Principles of frontal lobe function* (pp. 573-597). New York: Oxford University Press.

Knobe, J. (2003). Intentional action in folk psychology: an experimental investigation. *Philosophical Psychology, 16,* 309-324.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences, 9,* 355–7.

Kohlberg, L. (1969). Style and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research*. Chicago: Rand McNally.

Kohlberg, L. (1984). *The psychology of moral development.* San Francisco: Harper & Row.

Kortenkamp, K. V., & Moore, C. F. (2001). Ecocentrism and anthropocentrism: Moral reasoning about ecological commons dilemmas. *Journal of Environmental Psychology, 21,* 261-272.

Kortenkamp, K. V., & Moore, C. F. (2009). Children's moral evaluations of ecological damage: The effect of biocentric and anthropocentric intentions. *Journal of Applied Social Psychology, 39,* 1785-1806.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences, 110,* 5648-5653.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330,* 1830-1834.

Krause, J. (1993). The relationship between foraging and shoal position in a mixed shoal of roach (*Rutilus-Rutilus*) and chub (*Leuciscus-Cephalus*) – a field-study. *Oecologia, 93,* 356-359.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science, 14,* 402-408.

LaFraniere, P., & Charlesworth, W. R. (1983). Dominance, attention, and affiliation in a preschool group: A nine-month longitudinal study. *Ethology and Sociobiology, 4,* 55-67.

Laland, K. N., & Galef, B. G. (2009). *The question of animal culture.* Cambridge: Harvard University Press.

Lasker, E. (1907). *Struggle.* New York: Lasker's Publishing Company.

Lee, Y., Yun, J. E., Kim, E. Y., & Song, H. (2015). The development of infants' sensitivity to behavioral intentions when inferring others' social preferences. *PLoS ONE, 10,* e0135588.

Lepsien, J., & Nobre, A. C. (2006). Cognitive control of attention in the human brain: Insights from orienting attention to mental representations. *Brain Research, 1105,* 20-31.

Leslie, A. (1987). Pretense and representation in infancy: The origins of 'theory of mind'. *Psychological Review, 94,* 84-106.

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology, 50,* 45-85.

Leslie, A. M., Friedman,, O. & German, T. P. (2004). Core mechanisms in 'Theory of mind'. *Trends in Cognitive Sciences, 12,* 258-533.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: theory of mind and moral judgment. *Psychological Science, 5,* 421-427.

Leslie, A. M., Mallon, R., & DiCorcia, J. (2006). Transgressors, victims, and cry babies: Is basic moral judgment spared in autism? *Social Neuroscience, 1,* 270-283.

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science, 1,* 247-253.

Lewis, H. (1974). *Leaders and followers: Some anthropological perspectives.* Reading, MA: Addison-Wesley.

Li, J., Zhu, L., & Gummerum, M. (2014). The relationship between moral judgment and cooperation in children with high-functioning autism. *Scientific Reports, 4,* 4314.

Light, L. L. (1991). Memory and aging: Four hypotheses in search of data. *Annual Review of Psychology, 42,* 333–376.

Lourenco, S. F., Bonny, J. W., & Schwartz, B. L. (2016). Children and adults use physical size and numerical alliances in third-party judgments of dominance. *Frontiers in Psychology, 6,* 2050.

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *The British Journal of Developmental Psychology, 30,* 1-13.

Luo, Y. (2011). Do 10-month-old infants understand false beliefs? *Cognition, 121,* 289-298.

Lyn, H., Franks, B., & Savage-Rumbaugh, S. (2008). Precursors of morality in the use of the symbols ‘good’ and ‘bad’ in two bonobos (*Pan paniscus*) and a chimpanzee (*Pan troglodytes*). *Language & Communication, 28,* 213-224.

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

Manning, J. E. (2015, December 1). Membership of the 114th Congress: A Profile. Retrieved from http://www.senate.gov/CRSReports/crs-publish.cfm?pid=%260BL*RLC2%0A

Marcus Aurelius (II century AD/2006). *Meditations*. Penguin Classics.

Margoni, F., Baillargeon, R., & Surian, L. (2016). *Infants distinguish between dominance and leadership.* Manuscript submitted.

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

Margoni, F., & Surian, L. (2016a). Explaining the U-shaped development of intent-based moral judgments. *Frontiers in Psychology, 7,* 219.

Margoni, F., & Surian, L. (2016b). Mental state understanding and moral judgment in children with Autistic Spectrum Disorder. *Frontiers in Psychology, 7,* 1478.

Marlowe, F. W. (2005). Hunter-gatherers and human evolution. *Evolutionary Anthropology, 14,* 54-67.

Marsh, H., Stavropoulos, J., Nienhuis, T., & Legerstee, M. (2010). Six- and nine-month-old infants discriminate between goals despite similar action patterns. *Infancy, 15,* 94-106.

Mascaro, O., & Csibra, G. (2012). Representation of stable social dominance relations by human infants. *Proceedings of the National Academy of Sciences, 109,* 6862-6867.

Mascaro, O., & Csibra, G. (2014). Human infants' learning of social structures: The case of dominance hierarchy. *Psychological Science, 25,* 250-255.

Maslow, A. H. (1936). A theory of sexual behavior in infra-human primates. *Journal of Genetic Psychology, 48,* 310-336.

McGinley, M., & Carlo, G. (2007). Two sides of the same coin? The relations between prosocial and physically aggressive behaviors. *Journal of Youth and Adolescence, 36,* 337–349.

McNair, S., Okan, Y., Hadjichristidis, C., & Bruine de Bruin, W. (2016). *Aging effects in moral judgment: Older adults are more deontological than younger adults.* Manuscript submitted for publication.

Meristo, M., & Surian, L. (2013). Do infants detect indirect reciprocity? *Cognition, 129,* 102-113.

Meristo, M., & Surian, L. (2014). Infants distinguish antisocial actions directed towards fair and unfair agents. *PLoS ONE, 5,* e110553.

Milfont, T. L., & Duckitt, J. (2010). The environmental attitudes inventory: A valid and reliable measure to assess the structure of environmental attitudes. *Journal of Environmental Psychology, 30,* 80-94.

Milgram, S. (1974). *Obedience to authority: An experimental view.* New York: Harper & Row.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167-202.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49-100.

Moran, J. D., & O'Brien, G. (1983). The development of intention-based moral judgments in three- and four-year-old children. *The Journal of Genetic Psychology, 143,* 175-179.

Moran, J. M. (2013). Lifespan development: The effects of typical aging on theory of mind. *Behavioural Brain Research, 237,* 32–40.

Moran, J. M., Jolly, E., & Mitchell, J. P. (2012). Social-cognitive deficits in normal aging. *The Journal of Neuroscience, 32,* 5553–5561.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceeding of National Academy of Sciences, 108*, 2688-2692.

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology, 138,* 535-545.

Moriguchi, Y., & Hiraki, K. (2009). Neural origin of cognitive shifting in young children. *Proceedings of the National Academy of Sciences, 106,* 6017-6021.

Moriguchi, Y., & Hiraki, K. (2011). Longitudinal development of prefrontal function during early childhood. *Developmental Cognitive Neuroscience, 1*, 153-162.

Moriguchi, Y. (2014). The early development of executive function and its relation to social interaction: A brief review. *Frontiers in Psychology, 5,* 388.

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

Nelson, S. A. (1980). Factors influencing young children's use of motives and outcomes as moral criteria. *Child Development, 51*, 823-829.

Nichols, S. (2004). *Sentimental rules.* New York: Oxford University Press.

Nisan, M. (1987). Moral norms and social conventions: A cross-cultural comparison. *Developmental Psychology, 23,* 719-725.

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

Nucci, L. (1981). Conceptions of personal issues: A domain distinct from moral or societal concepts. *Child Development, 49,* 400-407.

Nucci, L. (1985). Social conflict and the development of children's moral and conventional concepts. *New Direction in Child Development, 29,* 55-70.

Nucci, L. (2001). *Education in the moral domain*. Cambridge: Cambridge University Press.

Nye, J. S. (2008). *The powers to lead.* New York: Oxford University Press.

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B Biological Sciences, 276,* 323-329.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false belief? *Science, 308,* 255-258.

Over, H., & Carpenter, M. (2009). Eighteen-month-old infants show increased helping following priming with affiliation. *Psychological Science, 20,* 1189-1193.

Overbeck, J. R. (2010). Concepts and historical perspectives on power. In A. Guinote and T. K. Vescio (Eds.), *The social psychology of power* (pp. 19-45). New York: Guildford Press.

Parten, M. B. (1933). Leadership among preschool children. *The Journal of Abnormal and Social Psychology, 27,* 430-440.

Patil, I., Melsbach, J., Hennig-Fast, K., & Silani, G. (2016). Divergent roles of autistic and alexithymic traits in utilitarian moral judgments in adults with autism. *Scientific Reports, 6,* 23637.

Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology, 5,* 501.

Pellizzoni, S., Siegal, M., & Surian, L. (2009). Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology, 45,* 289-295.

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences, 16,* 519-525.

Phillips, L. H., MacLean, R. D., & Allen, R. (2002). Age and the understanding of emotions: Neuropsychological and sociocognitive perspectives. *Journal of Gerontology: Psychological Sciences, 57B,* 526–530.

Piaget, J. (1932). *The moral judgment of the child*. London: Kegan Paul.

Pietraszewski, D., & Shaw, A. (2015). Not by strength alone: Children's conflict expectations follow the logic of the asymmetric war of attrition. *Human Nature, 26,* 44-72.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*, 653–660.

Poama, A. (2012). Dangerous instruments? Constructing risk and culpable drivers through the criminalization of negligence. *British Journal of Criminology, 52,* 932-952.

Pratt, M. W., Diessner, R., Pratt, A., Hunsberger, B., & Pancer, S. M. (1996). Moral and social reasoning and perspective taking in later life: A longitudinal study. *Psychology and Aging, 11,* 66-73.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Pun, A., Birch, S. A. J., & Baron, A. S. (2016). Infants use relative numerical group size to infer social dominance. *Proceedings of the National Academy of Sciences,* 10.1073/pnas.1514879113

Rai, T. S., & Fiske, A. P. (2011). Moral Psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118,* 57-75.

Rands, S. A., Cowlishaw, G., Pettifor, R. A., Rowcliffe, J. M., & Johnstone, R. A. (2003). Spontaneous emergence of leaders and followers in foraging pairs. *Nature, 423,* 432-434.

Reamer, F. G., & Racette, M. J. (2015). *Risk management in social work: Preventing professional malpractice, liability, and disciplinary action.* New York, NY: Columbia University Press.

Reynolds-Keefer, L., Johnson, R., Dickenson, T., & McFadden, L. (2009). Validity issues in the use of pictorial likert scales. *Studies in Learning Evaluation Innovation and Development, 6,* 15-24.

Reuter-Lorenz, P. A., & Sylvester, C. Y. (2005) The cognitive neuroscience of aging and working memory. In: Cabeza R, Nyberg L, Park D, editors. *The cognitive neuroscience of aging* (pp. 186-217). New York: Oxford University Press.

Rhine, R., Hill, S., & Wanderuff, S. (1967). Evaluative responses of preschool children. *Child Development, 38,* 1035-1042.

Richardson, C. B., Mulvey, K. L., & Killen, M. (2012). Extending social domain theory with a process-based account of moral judgments. *Human Development, 55,* 4-25.

Richerson, P. J., & Boyd, R. (2006). *Not by genes alone.* Chicago: University of Chicago Press.

Ridgeway, C. L., & Diekeman, D. (1989). Dominance and collective hierarchy formation in male and female task groups. *American Sociological Review, 54,* 79-93.

Rogé, B., & Mullet, E. (2011). Blame and forgiveness judgments among children, adolescents and adults with autism. *Autism,* 15, 702-712.

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O., & Convit, A. (2007). Who cares? Revisiting empathy in Asperger syndrome. *Journal of Autism and Developmental Disorders, 37,* 709-715.

Rogers, J., Viding, E., Blair, J., Frith, U., & Happé, F. (2006). Autism spectrum disorder and psychopathy: Shared cognitive underpinnings or double hit? *Psychological Medicine, 36,* 1789-1798.

Rosenbaum, T. (2004). *The myth of moral justice.* New York: HarperCollins.

Rosmini, A. (1840/1989). *Conscience*. Durham: Rosmini House.

Rousseau, J. J. (1762/1913). *Social contract and discourses.* Trans. D. H. Cole, New York: Dutton.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296-320.

Russell, B. (1938). *Power: A new social analysis.* New York: Routledge.

Sackin, S., & Thelen, E. (1984). An ethological study of peaceful associative outcomes to conflict in preschool children. *Child Development, 55,* 1098-1102.

Salthouse, T. A. (2004). What and when of cognitive aging. *Current Directions in Psychological Science, 13,* 140–144.

Salvano-Pardieu, V., Blanc, R., Combalbert, N., Pierratte, A., Manktelow, K., Maintier, C., et al. (2015). Judgment of blame in teenagers with Asperger's syndrome. *Thinking & Reasoning, 22,* 251-273.

Schleifer, M., Shultz, T. R., & Lefebvre-Pinard, M. (1983). Children's judgments of causality, responsibility and punishment in cases of harm due to omission. *British Journal of Developmental Psychology, 1,* 87-97.

Schlottmann, A., Surian, L., & Ray, E. (2009). Causal perception of action-and-reaction sequences in 8- to 10-month-old infants. *Journal of Experimental Child Psychology, 103,* 87-107.

Schmidt, M. F. H., & Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. *PLoS ONE, 6,* e23223.

Schöppner, B., Sodian, S., & Pauen, S. (2006). Encoding action roles in meaningful social interaction in the first year of life. *Infancy, 9,* 289-311.

Schultz, P. W. (2000). Empathizing with nature: The effects of perspective taking on concern for environmental issues. *Journal of Social Issues, 56,* 391-406.

Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. *Developmental Science, 15,* 181-193.

Serafin, M., & Surian, L. (2004). Il Test degli Occhi: Uno strumento per valutare la teoria della mente. *Giornale Italiano di Psicologia, 31,* 839–860.

Shantz, C. U. (1987). Conflicts between children. *Child Development, 58,* 283-305.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness.* New York, NY: Springer Verlag.

Sheskin, M., Chevallier, C., Lambert, S., & Baumard, S. (2014). Life-history theory explains childhood moral development. *Trends in Cognitive Sciences, 18,* 613-615.

Shulman, C., Guberman, A., Shiling, N., & Bauminger, N. (2012). Moral and social reasoning in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 42,* 1364-1376.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science, 13,* 238-253.

Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science, 17,* 97-108.

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

Shweder, R., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in younger children* (pp. 1-83). Chicago: University of Chicago Press.

Shweder, R., Much, N., Mahapatra, M., & Park, L. (1997). The “big three” of morality (autonomy, community, divinity) and the “big three” explanations of suffering. In A.

M. Brandt & P. Rozin (Eds.), *Morality and Health* (pp. 119-169). New York: Routledge.

Sidanius, J., & Pratto, F. (1999). *Social dominance theory: An intergroup theory of social hierarchy and oppression.* New York: Cambridge University Press.

Siegal, M. (1982). *Fairness in children. A social-cognitive approach to the study of moral development*. New York, NY: Academic Press.

Siegal, M., & Peterson, C. C. (1998). Preschoolers' understanding of lies and innocent and negligent mistakes. *Developmental Psychology, 34,* 332-342.

Silk, J. B. (2007). Social components of fitness in primate groups. *Science, 317,* 1347-1351.

Silver, M., & Oakes, P. (2001). Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others. *Autism, 5,* 299-316.

Singer, P. (2011). *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science, 23,* 196-204.

Smetana, J. G. (2006). Social-cognitive 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: Lawrence Erlbaum Associates.

Smetana, J., & Braeges, J. (1990). The development of toddlers' moral and conventional judgments. *Merrill Palmer Quarterly, 36,* 329-346.

Smetana, J., Jambon, M., Conry-Murray, C., & Sturge-Apple, M. (2012). Reciprocal associations between young children's developing moral judgments and theory of mind. *Developmental Psychology, 48,* 1144-1155.

Smith, A. (1759/1948). *A theory of moral sentiments*. New York: Hafner.

Snelgar, R. S. (2006). Egoistic, altruistic, and biospheric environmental concerns: Measurement and structure. *Journal of Environmental Psychology, 26,* 87-99.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18,* 587-592.

Steele, S., Joseph, R., & Tager-Flusberg, H. (2003). Brief report: Developmental change in theory of mind abilities in children with autism. *Journal of Autism and Developmental Disorders, 33,* 461-467.

Stein, N. L., & Glenn, G. C. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing.* Vol. 2. Norwood, N.J.: Ablex.

Stern, P. C., & Dietz, T. (1994). The value basis of environmental concern. *Journal of Social Issues, 50,* 65-84.

Strayer, F. F., & Strayer, J. (1976). An ethological analysis of social agonism and dominance relations among preschool children. *Child Development, 47,* 980-989.

Surber, C. F. (1977). Developmental processes in social inference: Averaging of intentions and consequences in moral judgment. *Developmental Psychology, 13,* 654-665.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18,* 580-586.

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology, 30,* 30-44.

Surian, L., & Leslie, A. (1999) Competence and performance in false belief understanding: A comparison of autistic and three-year-old children. *British Journal of Developmental Psychology, 17,* 131-145.

Svetlova, M., Nichols, S. R., & Brownell, C. A. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. *Child Development, 81,* 1814-1827.

Swanton, C. (2003). *Virtue Ethics: a Pluralistic View.* Oxford: Oxford University Press.

Sze, J., Gyurak, A., Goodkind, M. S., & Levenson, R. W. (2012). Greater prosocial and empathic responding in late life. *Emotion, 12,* 1129–1140.

Takeda, T., Kasai, K., & Kato, N. (2007). Moral judgment in high-functioning pervasive developmental disorders. *Psychiatry and Clinical Neurosciences, 61,* 407-414.

Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representing building blocks of active resource-transfer events in human infants. *Cognition, 137,* 47-62.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from implicit to an explicit understanding of false belief from infancy to preschool age. *The British Journal of Developmental Psychology, 30,* 172-187.

Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science, 331,* 477-480.

Thomson, J. (1985). The trolley problem. *Yale Law Journal, 94,* 1395–1415.

Tindall, R. C., & Ratliff, R. G. (1974). Interaction of reinforcement conditions and developmental level in a two-choice discrimination task with children. *Journal of Experimental Child Psychology, 18,* 183-189.

Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology, 44,* 187-194.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28,* 675-735.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of cooperation: The interdependence hypothesis. *Current Anthropology, 53,* 673-692.

Tooby, J., & Cosmides, L. (1992). Psychological foundation of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19-136). New York: Oxford University Press.

Turiel, E. (1978). Social regulations and domains of social concepts. In Damon W (Ed), *New directions for child development, vol. 1 Social Cognition*. San Francisco: Jossey-Bass.

Turiel, E. (1983). *The development of Social Knowledge: Morality and Convention.* Cambridge: Cambridge University Press.

Turiel, E. (2014). Morality, epistemology, development, and social opposition. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (2nd ed., pp. 3-22). New York: Psychology Press.

United Nation (2012). UNFPA report. Retrieved from https://www.unfpa.org/webdav/site/global/shared/documents/publications/2012/UNFPA-Report-Chapter1.pdf

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bullettin, 134,* 383-403.

Van Vugt, M. (2006). Evolutionary origins of leadership and followership. *Personality and Social Psychology Review, 10,* 354-371.

Vivanti, G., McCormick, C., Young, G. S., Abucayan, F., Hatt, N., Nadig, A., Ozonoff, S., & Rogers, S. J. (2011). Intact and impaired mechanisms of action understanding in autism. *Developmental Psychology, 47,* 841-856.

von Clausewitz, C. (1832/1984). *On war.* New Jersey: Princeton University Press.

von Rueden, C., Gurven, M., & Kaplan, H. (2011). Why do men seek status? Fitness payoffs to dominance and prestige. *Proceedings of the Royal Society B: Biological Sciences, 278,* 2223.

von Rueden, C., & Van Vugt, M. (2015). Leadership in small-scale societies: Some implications for theory, research, and practice. *The Leadership Quarterly, 26,* 978-990.

Wainryb, C., Brehl, B., & Matwin, S. (2005). Being hurt and hurting others: Children's narrative accounts and moral judgments of their own interpersonal conflicts. *Monographs of the Society for Research in Child Development, 70* (3, Serial No. 281).

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science, 311,* 1301-1303.

Weber, M. (1919/1994). Politics as a Vocation. In Lassman P. & Speir R. (eds.), *Max Weber, Political Writings,* Cambridge University Press.

Weber, M. (1946). Class, status, party. In H. H. Gerth & C. Wright Mills (Eds.), *Max Weber: Essays in sociology* (pp. 180-195). Oxford, UK: Routledge.

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

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

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,* 655-684.

White, L. (1967). The historical roots of our ecological crisis. *Science, 155,* 1203-1207.

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

Williams, G. L. (1953). *Criminal law: The general part.* London: Stevens & Sons.

Wilson, E. O. (2012). *The social conquest of earth.* New York: Norton.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103-128.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69,* 1-34.

World Health Organization. (2015, September). Ageing and health. Retrieved from http://www.who.int/mediacentre/factsheets/fs404/en/

Wright, J. C., Zakriski, A. L., & Fisher, P. (1996). Age differences in the correlates of perceived dominance. *Social Development, 5,* 24-40.

Wynn, K. (2008). Some innate foundations of social and moral cognition. In P. Carruthers, S. L. Laurence, & S. Stich (Eds.), *The innate mind: Volume 3. Foundations and the future* (pp. 330-347). Oxford: Oxford University Press.

Yirmiya, N., Sigman, M. D., Kasari, C., & Mundy, P. (1992). Empathy and cognition in high-functioning children with autism. *Child Development, 63,* 150-160.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of National Academy of Sciences, 107,* 6753-6758.

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,* 8235-8240.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology, 1,* 333-349.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47,* 2065-2072.

Young, L., Scholz, J., & Saxe, R. (2011). Neural evidence for 'intuitive prosecution': The use of mental state information for negative moral verdicts. *Social Neuroscience, 6,* 302-315.

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass, 7*, 585–604.

Yuill, N. (1984). Young children's coordination of motive and outcome in judgments of satisfaction and morality. *British Journal of Developmental Psychology, 2,* 73-81.

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

Zalla, T., Barlassina, L., Buon, M., & Leboyer, M. (2011). Moral judgment in adults with autism spectrum disorders. *Cognition, 121*, 115-126.

Zalla, T., & Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophy and Psychology*, *2,* 681-698.

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

Zmyj, N., Buttelmann, D., Carpenter, M., & Daum, M. (2010). The reliability of a model influences 14-month-olds' imitation. *Journal of Experimental Child Psychology, 106,* 208-220.