November 2016

Infants' reasoning about agents' identity: the case of sociomoral kinds

Hernando Taborda
University of Massachusetts Amherst

INFANTS' REASONING ABOUT AGENTS' IDENTITY: THE CASE OF SOCIOMORAL KINDS

A Dissertation Presented

by

HERNANDO TABORDA OSORIO

Submitted to the Graduate School of the
University of Massachusetts, Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2016

Psychological and Brain Sciences

INFANTS' REASONING ABOUT AGENTS' IDENTITY: THE CASE OF SOCIOMORAL KINDS

A Dissertation Presented

by

HERNANDO TABORDA OSORIO

Approved as to style and content by:

_____
Erik Cheries, Chair

_____
Neil Berthier, Member

_____
Ronnie Janoff-Bulman, Member

_____
Louise Antony, Member

_____
Harold D. Grotevant, Chair of Department
Department of Psychological and Brain Sciences

ABSTRACT

INFANTS' REASONING ABOUT AGENTS' IDENTITY: THE CASE OF SOCIOMORAL KINDS

SEPTEMBER 2016

HERNANDO TABORDA OSORIO, B.S., UNIVERSIDAD NACIONAL DE COLOMBIA

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

PH.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Erik W. Cheries, Ph.D.

Recent studies in development psychology suggest that early on infants are able to distinguish characters who display a cooperative behavior from characters who display an antisocial behavior. The current research builds on these findings and aims at determining the extent to which infants possess the sociomoral distinction of "good" and "mean" agents. In particular, we propose that infants represent sociomoral behaviors through kind-based categories. This hypothesis was tested in the current research across 5 different experiments by investigating how infants represent the identity of agents in sociomoral situations. Experiment 1 used a looking-time paradigm to demonstrate 11-month-old infants' bias to individuate distinct agents based upon their "mean" and "nice" behaviors in a spatiotemporal ambiguous situation. Experiment 2 and 3 ruled out alternative explanations of this effect by controlling for the number of actions presented and differences in motion, respectively. These findings suggest that infants expect agents to display coherent sociomoral behaviors over time in a particular context. Experiment 4 tested whether infants' are biased to identify prosocial agents more by their internal than their external properties. Fourteen-month-olds showed a bias to identify the 'helper' character based

on the color of its internal properties. Experiment 5 aimed to clarify whether infants have this biased because they attribute a causal role to the agents' internal properties. In two different conditions the causal relevance of the agents' internal or external property was manipulated. We hypothesized that when the causal relevance of the internal property was undermined infants would no longer be biased toward the agents' internal properties when identifying it as the "helper" character. So far, the results do not show a clear support for this hypothesis. Overall, the results of all these experiments indicate that infants represent sociomoral behaviors in a relatively categorical fashion, and more strongly associated to the agents' internal rather than external properties. These findings are discussed from both a strong version of kind-based representations in terms of intrinsic natural kinds, and a weaker version in terms of more graded and extrinsic sociomoral kinds.

Keywords: agents, development, infants, kind categories, sociomoral dispositions.

TABLE OF CONTENTS

Page

ABSTRACT……...iv

LIST OF FIGURES……...……...viii

CHAPTER

1. INTRODUCTION…….……..1

2. LITERATURE REVIEW…….……...5

2.1. The concept of "agency" in infancy.…….…….……5
2.2. Kind concepts…….……..12
2.3. Development of Sociomoral Concepts…….…….…….……..19

3. THE PRESENT RESEARCH.…… 27

4. STUDY 1: INDIVIDUATION OF AGENTS BY MORAL DISPOSITIONS…….29

4.1. Experiment 1…….……29
4.1.1. Introduction……..……..29
4.1.2. Method……...32
4.1.2.1. Participants…….……32
4.1.2.2. Materials……...…….33
4.1.2.3. Design and Procedure……...….34
4.1.3. Results and Discussion……..36
4.2. Experiment 2…….……38
4.2.1. Method……...……38
4.2.1.1. Participants……...……...38
4.2.1.2. Materials, Design, and Procedure……..38
4.2.2. Results and Discussion……..39
4.3. Experiment 3……..…41
4.3.1. Method……...……41
4.3.1.1. Participants…….41
4.3.1.2. Materials, Design, and Procedure……..42
4.3.2. Results and Discussion…….…….42
4.4. General Discussion…….……..44

5. STUDY 2: INSIDES AND MORAL DISPOSITIONS…….46

5.1. Experiment 1…….……46
5.1.1. Introduction……..…..46
5.1.2. Method……..…….50

5.1.2.1. Participants……50
5.1.2.2. Apparatus……50
5.1.2.3. Procedure……51
5.1.3. Results and Discussion……52
5.2. Experiment 2……55
5.2.1. Method……55
5.2.1.1. Participants……55
5.2.1.2. Apparatus……55
5.2.1.3. Procedure……55
5.2.2. Results and Discussion……56

6. GENERAL DISCUSSION……58

APPENDICES

A. STIMULI PICTURES STUDY 1……66

B. STIMULI PICTURES STUDY 2……68

BIBLIOGRAPHY……70

LIST OF FIGURES

Figure Page

1. Mean Looking-Time Results Experiment 1 ……...38

2. Mean Looking-Time Results Experiment 2 ……41

3. Mean Looking-Time Results Experiment 3 ……44

CHAPTER 1

INTRODUCTION

During the last 30 years extensive research in infant cognition has unveiled some key aspects about how infants keep track of an object's identity over time. Indeed, a vast literature suggests that object individuation abilities—judging if an object is the same or a different instance of one seen before—seems to be a fundamental cognitive skill already present in early stages of development (for reviews see, Carey & Xu, 2001; Xu, 2007). For example, by as early as 2 months of age infants are able to determine if they are observing one or two objects in an event based upon evidence that the objects occupy distinct spatial locations across time (Aguiar & Baillergeon, 1999; Van de Walle, Carey & Prevor, 2000; Xu & Carey, 1996).

The debate in infant cognition about whether behavioral performance reflects mere perceptual abilities or both perceptual and high-level conceptual knowledge (e.g., Carey & Spelke, 1994) also arises in the literature on object individuation. Some research advocates for a low-level interpretation of infants' individuation abilities by invoking automatic visuospatial attentional mechanisms (Scholl & Leslie, 1999), while other researchers emphasize the early emergence of more conceptual-based mechanisms (Carey & Xu, 2001). In order to better understand the mechanisms involved in the process of object individuation researchers have examined the type of properties infants use to individuate objects that belong to different ontological and taxonomical categories (Baillargeon, Stavans, Wu, Gertner, Setoh, Kittredge, & Bernard, 2012). In this regard, the ontological distinction between agents and non-agents has been one important domain where both low- and high-level explanations have been empirically contrasted. For

instance, 10-month-olds have been shown to individuate agents and inanimate objects, but fail to individuate two perceptually dissimilar agents (Surian & Caldi, 2010), suggesting that at least in some circumstances infants are able to recruit abstract conceptual knowledge – such as a distinction between animates and inanimates—to keep track of the objects' identity over time

In addition to looking at how infants individuate agents from non-agents, recent research has examined the influence of conceptual knowledge in representing the identity of individual agents. In particular, previous studies with older children have shown that intuitive biological knowledge constrains the way they represent the identity of living beings. For example, 7-year-old children believe that an animals' internal properties are more likely than external properties to determine their categorical membership (e.g. whether it is a cat or a dog; Gelman, 2003; Keil, 1989) as well as their individual identity to the extent that each individual is represented as belonging to only one immutable category (Gutheil & Rosengren, 1996). A similar reasoning bias about the relative importance of internal and external properties in the representation of agents' identity has been shown in infants. When 13-month-olds observe a transparent object with contrasting internal and external properties they tend to use the color of internal properties as a more reliable cue of identity than its external properties (Taborda-Osorio & Cheries, 2015). This result suggests that from early on domain-specific biological knowledge seems to constrain the way infants reason about agent identity.

The current set of proposed studies further explores the unique ways in which infants represent the identity of agents. Although using internal properties could be relevant when reasoning about animals embedded in taxonomical categories (Atran,

1998; 1999), adults tend to use psychological characteristics as a better indicator of people's identity (Haslam, Bastian, & Bissett, 2004). Among all psychological characteristics that people use, sociomoral dispositions is one of the most elemental ways that even preschool-aged children use to distinguish people, and it is relevant when representing peoples' identity. For example, children use labels like "mean" and "nice" to refer to people and to infer their future behaviors, mental states and emotions (Heyman & Gelman, 1998). Children's explanations about the source of these properties transitions from an early belief that being "mean" or "nice" is attributed to one's nature to a point where nurture is viewed as the main source of one's sociomoral attributes (Heyman & Gelman, 2000). This pattern of results with children has led some researchers to suggest that children represent sociomoral categories like intrinsic and essentialized attributes, meaning that they are immutable over time, and independent from external features and peoples' beliefs. In other words, being mean or nice is represented like an essential psychological attribute that defines people's identity to some extent. This type of intrinsic and essentialized category is referred in the psychological literature as a *kind concept*.

Previous developmental studies have demonstrated that as early as 6 months of age infants are able to distinguish agents based on their sociomoral dispositions, such as "mean" and "nice" or "helper" and "hinderer" (for a review see Hamlin, 2013). The main goal of the current set of studies is to determine to which extent infants represent those sociomoral dispositions like kind concepts, and hence like an identity-determining attribute of some agentive entities. The main motivation of studying infants is to give some insight into the developmental origins of moral concepts. In concrete terms, we want to test infants between 11 and 14 months of age because this is the age range at

which language production starts. This is crucial because unlike concepts in other domains (e.g. natural kinds) social categories appear to be strongly influenced by cultural factors and language. Therefore, demonstrating that moral categories are represented like kind concepts at the onset of language acquisition would suggest the presence of abstract initial underpinnings that may guide subsequent learning of moral knowledge.

The current proposal is divided into five sections. First, the literature review, where the concepts of "agent", "kind concepts" and "social kinds" are further developed. Second, the overall description of the current research. In the third and fourth sections Study 1, "Individuation of Agents by Moral Dispositions" and Study 2, "Insides and Moral Dispositions" are presented. Finally, in the last section some preliminary conclusions will be presented.

CHAPTER 2

LITERATURE REVIEW

2.1. The Concept of "Agency" in Infancy

For adults, the conceptual distinction between agents and non-agents relies mostly on the attribution of mental states, such as ascribing goals and dispositions to agentive entities in order to explain their behavior. On the other hand, the use of pure mechanical laws seems to be enough to explain the behavior of non-agentive, or inanimate, entities (Carey, 2009). Children's understanding of this ontological difference emerges very early on in life. For example, 5-month-old infants know that inert objects move together if and only if they touch (principle of contact), however, they also believe that human action is not constrained by this principle and are not surprised if they witness two human-like shapes moving without contacting each other (Spelke, Phillips, & Woodward, 1995). Likewise, 5-month-old infants also expect that inert objects but not people follow a continuous spatiotemporal path when moving across the stage (Kuhlmeier, Bloom & Wynn, 2004). Possibly, they believe that humans are somehow special entities that do not necessarily follow the same physical rules of inanimate objects. Other studies have also demonstrated that the conceptual distinction between agents and inanimate objects has consequences in the way infants categorize (Mandler, 2004) and individuate objects (Surian & Caldi, 2010).

This initial distinction between agents and non-agents leads to the following important question about the infants' understanding of agents: What type of properties do infants use to *identify* agents as different from non-agents? There are two classic answers

to this question in the developmental literature. The first claims that the presence of featural cues resembling a human being are necessary for identifying agents. The second claims that exhibiting certain motion cues, such as self-propelled behavior, is enough to trigger a representation of agency.

Woodward has been probably the most influential advocate for the first answer, in a seminal study, Woodmard (1998) tested 9-month-old infants' ability to encode actions like goal-oriented behaviors in two different conditions; infants were either habituated to repeated presentations of a hand reaching one of two objects placed on a stage or they witnessed a rod touching one of the same two objects. Then, the objects' positions were switched to see if infants at this age expect the hand and the rod to be oriented towards the same object, regardless of its current position or if they expect to see them following the same spatiotemporal trajectory. The results of this study revealed that infants who participated in the hand condition expected reaching towards the same object, whereas infants in the rod condition were relatively insensitive to the object and instead reacted to the rod changing its spatiotemporal trajectory on the stage. Subsequent studies using eye tracking demonstrated that infants not only are surprised when the hand reaches a different object but they also actively anticipate the hand's motion once the objects switch places (Cannon & Woodward, 2012; Woodward & Cannon, 2013). Infants have also been shown to be able to use other human typical cues in addition to hands to infer intentionality, such as peoples' looking and pointing (Johnson, Slaughter & Carey, 1998; Tomasello, Carpenter, & Liszkowski, 2007).

According to Woodward (2009), these studies suggest that the infants' daily experience with hands reaching towards objects in a goal-oriented fashion may support

their ability to understand that hand-shaped objects but not rods or other inanimate objects display agentive behavior. Additional support for this hypothesis comes from Guajardo and Woodward (2004) where 7- and 9-month-old infants were shown the same experimental setup as in Woodward (1998) but the hands' surface properties were obscured by a glove, resulting in the infants' inability to interpret hands' motion as goal-oriented behavior. Woodward (2009) also argues that the infants' own experience with reaching behaviors rather than the observation of others engaged in reaching actions is the main source of their goal attribution abilities. For instance, 3-month-old infants who typically fail to interpret another person's hand movements as goal-oriented will do so successfully after they have been outfitted with Velcro-covered mittens that allow them to experience "grabbing" objects for themselves.

The second classic answer to the question of how infants identify agents was proposed by Premack. In his classic paper Premack (1990) proposed that self-propelled motion is the main cue by which infants distinguish agents from non-agents, in such a way that whenever an object displays self-propelled motion infants are predisposed to interpret intentional movements. This theory, however, has not received good empirical support. For example, in a study carried out by Csibra (2008), 6-month-old infants observed an inanimate object (a box) displaying self-propelled motion on a computer generated 3D stage. The box moved around an obstacle to reach a target in two different conditions: the box reached its target displaying either the same path across trials (single route condition) or different paths (variable route condition). The results of this experiment revealed that infants this age infer agency only in the variable condition and do not seem to treat self-propelled motion as a critical component for identifying agents.

Shimizu and Johnson (2004), and Luo and Baillargeon (2005) discovered other behavioral cues that when present in the experimental setup are enough to trigger the inference of agency. In these studies, infants observed an inanimate object engage in a choice task that resembles the original Woodward (1998) procedure. These researchers discovered that variables such as the presence of more than one object on the stage in a choice situation, and cues of attentional orientation (the box "looking" sequentially to both objects) are important for the attribution of agency.

The results of the previous studies are also in line with some theories that propose an early emergence of a conceptual distinction between mechanical agency and intentional agency (Carey, 2009; Leslie, 1994). For example, Leslie (1994) proposed that a domain-specific mechanism (ToBy) is devoted to analyze an object's physical and mechanical properties. As part of this analysis, the system uses the cue of self-propelled movement to represent an agent like an object with internal and renewable sources of energy. Therefore, this notion of agency is purely physical and has nothing to do with intentional behavior. As evidence for the concept of mechanical agency in the infants' mind, Leslie (1994) notices that when infants are exposed to videos of launching events (a ball hitting other ball) they assign different roles to the object that causes the motion and the object that receives the motion (agent and non-agent). Thus, when infants observe the same video played backwards they find it strange that the object that previously caused the motion is now the recipient of the movement (Leslie, 1988). According to Leslie (1994), the understanding of these launching events involves the distinction of agents and non-agents in mechanical terms, without referring to psychological constructs to make sense of objects' motion.

Although the previous studies present substantial evidence against the hypothesis that self-propulsion is a necessary cue for intentional agency, these same studies also seem to call into question the experience and featural account of agency proposed by Woodward insofar as in all these cases infants attribute agency to inanimate objects. Infants seem not to struggle with attributing goals to boxes on the stage. This suggests that neither self-propulsion nor featural information alone is necessary for infants to identify agents. Therefore, some researchers (Csibra and Gergely, 1998; Biro and Leslie, 2007; Baillargeon, Wu, Yuan, Li, & Luo, 2009; Biro, Csibra & Gergely, 2007; Luo and Choi, 2013) have proposed that the identification of agents is rooted in a specialized system of psychological reasoning that makes use of patterns of object behavior that gives evidence of internal control and perception of the environment. In other words, the object should appear as being context-sensitive and possessing free will.

Csibra and Gergely (1998), and Gergely and Csibra (2003) also point out that the representation of internal control not only depends on the perception of the agent's behavior itself but also on the perception of the physical context and the end state of that behavior. According to these authors, the representation of agency emerges only when all these three variables satisfy the principle of rational action, meaning that the object's behavior should be perceived like efficient in attaining an end state (or goal) under the restrictions of the current physical context. Efficiency may be judged as the most direct or the least effortful action to attain a goal. Evidence for this theory comes from several studies where infants are shown inanimate objects displaying either a rational or an irrational action on the stage (Csibra, Biro, Koos, & Gergely, 2003; Csibra & Gergely, 2007; Csibra, Gergely, Biro, Koos, Brockbank, 1999). For example, in the classic study

of Gergely, Nadasdy, Csibra and Biro (1995), 12-month-old infants were shown an inanimate object (a ball) jumping over a barrier to reach a target (another ball), when in the test trials infants were shown the same scenario without the barrier, they expected the object to follow a straight trajectory to reach the target rather than the previous curve trajectory. Authors interpreted these results as demonstrating that infants are able to analyze the agents' behavior (the curved trajectory), the physical constraint (the barrier) and the end state (reach the target) together to infer what the most rational or efficient action the agent will display in the future. Subsequent studies demonstrated that infants are also able to use this tripartite representation in a productive way. For instance, infants are able to infer the agent's goal by taking into account the agent's behavior and the perceived physical constrains (Csibra, et. al., 2003).

Even though agentive entities have been distinguished from non-agentive entities based only on the attribution of mental states, it is unclear whether or not infants also attribute biological properties to agentive entities. Some researchers (Mandler, 2004; Carey, 1994) have provided a negative answer to this question by arguing that biological knowledge is the result of conceptual change along development. For instance, Carey (1994) proposes that only around 10 years of age do children display a biological understanding of typical biological process such as growth and respiration in terms of internal physiological mechanisms. Other researchers, however, have called into question this conclusion and have suggested that some rudimentary biological knowledge may be in place as early as in the first year of life (Gelman, 1990). Recent work has addressed this by testing whether or not 8-month-old infants attribute biological properties to agentive entities (Setoh, Wu, Baillargeon and Gelman (2013). Specifically, they tested

whether infants believe that self-propelled and agentive objects have something inside that could be responsible for the agent's motion. Setoh et. al (2013) tested this hypothesis by presenting objects on the stage in three different conditions across experiments: in the first condition the object was self-propelled and agentive, displaying features such as fur or contingent reaction, in the other two conditions objects displayed only one of the two properties alternatively. Once infants were familiarized to the object's motion on the stage the experimenter lifted the object showing it either filled or empty. The pattern of results revealed that infants expected the object to be filled only in the first condition when they had witnessed a self-propelled and agentive object. This study suggests first that from very early on infants distinguish biological agents from non-biological agents in terms of the presence of internal properties, and second that both self-propulsion and the display of agentive features (such as eyes or fur) seem to be necessary cues for the representation of biological agents.

In summary, infants seem to possess several specialized systems for representing and reasoning about agents. An innate physical system allows them to represent agents as possessing an internal source of energy, by detecting self-propelled motion as the main cue of agency. A psychological system allows infants to represent agents with intentional states, such as goals and dispositions, as a means to explain their behavior. Infants are able to use both featural information and patterns of behavior indicating context-sensitivity as the main cues of psychological agency. Finally, recent evidence suggest that infants may possess a biological reasoning system that allows them to represent some agents as having internal biological properties. The distinction between biological and

non-biological agents seems to rely on the simultaneous perception of self-propelled motion and agentive featural cues.

2.2. Kind Concepts

A great deal of work in the literature on kind concepts has focused its attention on how people represent and reason about natural kinds, including mainly animals and chemical compounds such as gold and water (Estes, 2003; Malt, 1994; Rips, 2011). A natural kind concept refers to a complex representational structure where multiple items are grouped together based on a non-obvious similarity, or essential property (Kripke, 1980; Putnam, 1975; Quine, 1969). In contrast to human made objects, natural kinds are supposed to be *objective* and *intrinsic* categories (Rips, 2011), meaning that what the essential property is does not depend on either human beliefs or the interactions with other objects in the world. Therefore, natural kinds are ahistorical categories which essence should be discovered through scientific research. Some philosophers and scientists have called into question both the search for essential properties as the main goal of a scientific enterprise and the existence itself of real essences in the world (Ereshefsky, 2010; Wilson, Barker & Brigandt, 2007). However, an essence is not necessarily something real, but rather an assumed property of some objects that may determine the referent of a word and serve to represent the ultimate cause of a pattern of observed correlations in an entity.

Although there has been some debate about whether or not people conceptualize the essential and hidden properties of a category as the main referent of a word (Braisby, Franks & Hampton, 1996; Malt, 1994), some evidence indicates that this may be the case

(Haukioja, 2014; Jylkka, 2008). For example, when adults are told a fictitious story where it has been discovered that the real chemical composition of water is XYZ rather than H2O, people tend to consider that what has been called water so far is not water any more but some other thing, despite having the same observable properties (Jylkka, Railo & Haukioja, 2009). Therefore, people seem to believe that the referent of a word depends on an external non-obvious property.

Similar to word definitions, there has been some debate about the causal understanding of essential properties. Some researchers deny that this could be the case insofar as people normally lack any specific knowledge about what the essential property of a category is (Strevens, 2000). However, in a classic paper Medin and Ortony (1989) propose that peoples' knowledge about those essential properties can be reduced to an overall assumption without specific content. Hence, in their model the essential property plays the role of a causal placeholder in the peoples' representation of natural kinds.

The assumed sharing of an essential property with causal potency across individuals in the same kind category allows people to engage in high order reasoning in different cognitive tasks, such as when categorizing objects (Medin & Ortony, 1989), keeping track of individuals over time (Wiggins, 1980), and making inductive inferences (Rips, 2004). Thus, people tend to outweigh hidden properties over observable features when categorizing animals. For example, people categorize mice and whales in the same group as mammals, although they look very different (Sloman, 2005). People also know that object identity can be preserved across some changes in external properties (e.g. a mouse changing color), but they are more reluctant to accept that object identity is preserved across a transformation in category membership (e.g. a mouse changing into a

whale without evidence of spatiotemporal continuity). In other words, categorical membership is represented as more central to the individual identity than superficial features (Hirsch, 1982). Additionally, knowing the category membership of an animal allows people to generalize properties to other animals in the same category. For instance, people readily transfer the property "warm blood" from a mouse to a whale but not from a mouse to a salmon. All three types of reasoning about natural kinds – categorization, individuation, and inductive inference- are closely related each other and depend on the assumption of a hidden essential property, in such a way that the possession of that essence determines both the categorical and the individual identity.

The process of categorization of natural kinds has been one of the most fertile grounds in psychology for exploring the hypothesis of essentialized categories. In a seminal study, Ahn (1998) discovered the *causal status effect* in categorization tasks where causal features are more important to category membership than less causal or effect features. For example, people consider that the feature *wings* in a bird is more important than the feature *fly* because flying is the consequence of having wings (Ahn, Gelman, Amsterlaw, Hohesnstein and Kalish, 2000; Ahn, Kalish, Gelman, Medin, Luhmann, Atran, Coley & Shafto, 2001; Ahn, Kim, Lassaline & Dennis, 2000; Kim & Ahn, 2002). According to Ahn et al., (2000), this effect explains why the assumption of an essential property as the original cause is more important in categorization than visible properties. Building on these results and other similar effects of causal knowledge, Rehder (2003), Rehder and Kim (2009), and Hayes and Rehder (2012) propose that categorization of natural kinds is a two-step process of diagnostic reasoning. The first step is to infer the presence of an unobservable feature from observable features, and the

second step is to determine the category membership from the unobservable features. In support of this theory, Rehder and Kim (2009) found that people assess visible features as more diagnostic of category membership when they are causally connected to unobservable features than when they do not. For example, participants were told that in Kehoe ants, blood high in iron sulfate (the underlying feature) was responsible for hyperactive immune system (observable features), while in Argentine ants, blood high in metallic sodium was merely correlated with fast digestion. When an animal exhibits both observable features people tended to categorize it as a Kehoe ant, presumably because the hyperactive immune system is a reliably diagnostic feature of a causal underlying property. These studies demonstrate that the assumption of an essential property is not just a way of representing the referent of natural kind concepts, but also has real implications in the way people reason about the world.

Several authors have suggested that the assumption of essentialized categories is proper only of natural kinds (Malt & Sloman, 2007; Sloman, 2005; Sloman & Malt, 2003). However, a growing body of research indicates that people may represent artifacts and diverse social categories in a similar fashion as natural kinds. In the case of artifacts, several authors (Bloom, 1996, 1998; Keil, Greif & Kerner, 2007; Matan & Carey, 2001; Rips, 1989) have suggested that similarly to biological entities artifacts such as chairs and boats are represented as entities that successfully express the intention of a designer to create a member of a particular artifact kind in a historical setting. Thereby, the way people differentiate the concept boat from the concept chair is by referring to the designer's intention to create an object that belongs to either the category of chairs or the category of boats. In this regard, the designer's intention is a non-obvious property that

groups together potentially dissimilar objects. Although function is an important property to represent artifacts, it is not an individuating factor because it is neither sufficient nor necessary to differentiate artifacts. For instance, people judge an object with a typical appearance of a boat but used like an off-shore jail to be a boat (Malt & Johnson, 1992), presumably because having the typical appearance of a boat indicates that the original creator's intention was to design a bout and not an off-shore jail. Evidence for the "design stance" of artifact categories has been found in different cultural contexts (Barrett, Laurence & Margolis, 2008).

Like biological and artifact kinds, it has been demonstrated that social categories are represented around a core of non-obvious properties responsible for the categorical identity of members in those categories (Haslam, 1998; Haslam & Ernst, 2002; Haslam, Rothschild & Ernst, 2000; Prentice & Miller, 2006; Prentice & Miller, 2007). For example, people tend to represent some social categories (e.g. gender) like highly uniform, discrete, immutable and objective. It has also been shown that adults represent some personality traits (e.g. intelligence, talkative and creative) as deep psychological characteristics, relatively immutable and highly informative about future people's behaviors, emotions and other mental states (Gelman, 2003; Haslam, Bastian & Bissett, 2004).

In a recent study, Ahn, Taylor, Kato, Marsh & Bloom (2013), demonstrated that people do not just represent the hidden essential properties as central in the structure of kind concepts, but also assume that they should be causally responsible for the observable properties. In one experiment adults were asked to rate how likely different members in four different types of categories –living things, artifacts, mental illness and medical

disorder- may share something that causes the typical features display in those members. When categories were named like kinds (e.g. by indicating that an instance belongs to a superordinate kind) participants tended to attribute a common cause, but when they were named like arbitrary groups participants were significantly less likely to attribute a common cause. This study demonstrates first a causal connection between a non-obvious property and visible properties in people's representation of kinds, and second that this representation seems to be common for all objects that are described as members of a kind, regardless the specific category.

Other studies found support for the hypothesis of a common representational structure of kind concepts by showing differences in the way people reason about kinds and non-kinds (Prasada, Hennefield, and Otap, 2012; Rips, 1989). In this line of research, Prasada, et, al. (2012) propose the "Distinct Representation Hypothesis", according to which the human conceptual system has two different ways to represent categories: kind representations and class representations. Kinds are understood to be intrinsically general, and supporting kind specifications where a category as a whole is represented as a member of a superordinate category. Thereby, a kind is represented as a specific way to realize the superordinate category. For example, a "dog" is a *kind of* animal, and a "sailboat" is a *kind of* boat. By contrast, a class representation is an arbitrary category where the group as a whole is not represented as a member of a superordinate category. Thus, "white bears" is not a kind of bear, and "blue buses" is not a kind of vehicle.

Even though it is possible to outline several similarities across different kind categories that suggest a common representational structure, some researchers point out important differences between people's representation of natural kinds and other kind

categories (Estes, 2003; Kalish, 1995, 2002). The most notable difference is that artifacts and social categories are not as essentialized as natural kinds. Thus, adults believe that animals of the same species share a real essence inside each exemplar, render them more resistant to changes in external appearance in categorization judgements. A radical example of this believe is the "genetic essentialism", according to which genes strongly determine the kind membership of an animal (Dar-Nimrod & Heine, 2011). By contrast, artifacts seem to be devoid of internal essences (e.g. nothing inside a hammer makes it the artifact it is), and are more conventionalized. As a consequence, membership in a natural kind is absolute (i.e. all or none), while membership in an artifact kind tend to be more graded (Estes, 2003). Similarly, some studies carried out by Haslam and colleagues demonstrate that not all social categories are equally essentialized (Haslam, et. al., 2000). Some of them (e.g. gender, race, and age) are represented like natural kinds or pseudo-natural kinds (Boyer, 1993), while others (e.g. politic affiliation and religion) are represented as possessing an underlying reality but less immutable over time.

A second important difference in category structure is that observable features are not equally diagnostic of category membership across different kinds. Thus, some studies conducted by Keil (1995) demonstrate that the perceived importance of properties in categorization judgements varies as a function of the type of kind. For example, changes in color are very important for categorizing chemical compounds but not for artifacts and living things, while changes in shape shows the reverse pattern. These results indicate that the relationship between observable and unobservable properties is not a simple one, and it may be supported by abstract theoretical beliefs about how properties in different conceptual domains are interconnected each other.

In summary, the representation of kind concepts in adults exhibits both unity and diversity. Unity in their basic organization where non-obvious properties are represented as causal-explanatory features, and in the role that kinds play in the human inferential system, supporting categorization, individuation and inductive generalizations. Diversity in domain-specific differences regarding the type of visible properties that are more diagnostic of membership, and in the degree that those categories are essentialized. Thus, for natural kinds the essential property is projected as part of the object's internal structure, while for artifacts the essential property is extrinsic and ultimately relies on the creator's mind. The questions to address now have to do with how this type of representation originates and develops over time. Although considerable progress has been made in how children understand several classes of natural kind concepts, the next section will focus mostly on children's understanding of biological concepts and artifacts.

2.3. Development of Sociomoral Concepts

Like biological kinds, children have been shown to essentialize social categories (Birnbaum, Deeb, Segall, Eliyahu & Diesendruck, 2010). Thus, Hirschfeld (1995, 1996) demonstrated that children as young as 4 years of age represent race as a more identity relevant property than other biological properties (e.g. body build). They also believe that racial identity is inherited from parents to children and maintain throughout life regardless the cultural context were the child is raised. For example, using a "switched at birth" task, children were told a story about two racially different couples who accidentally change their babies at birth. Then, they were shown two pictures of two school-age children, one black and the other white, and asked to choose which the correct

couple's baby was years later. Results show that 4 and 5-year-olds tend to choose based on racial correspondence, meaning that they represent race as an essential property transmitted through biological mechanisms (e.g. birth).

Similar conclusions have been reached with other social categories. Using the switched at birth task, Taylor (1996), and Taylor, Rhodes and Gelman (2009), demonstrated that until 9-10 years of age children hold the belief that gender-stereotypical properties are inherited and biologically transmitted. Following the studies of Gil-White (2001) about ethnicity with Mongolian communities, Birnbaum, et al. (2010) showed that Israeli children essentialize ethnic categories (e.g. Arabic and Jewish) by using inductive potential tasks. For example, when children are told that two different people share the same ethnic membership, they generalize psychological properties across both members. These results indicate that Israeli children represent ethnicity as a deep causal-explanatory feature. Language is another social category that has also been shown to trigger essentialist beliefs in children. Thus, Kinzler and Dautel (2012) demonstrated that 5-6-year-old children believe that the type of language (e.g. French or English) spoken by a person but not race will remain stable throughout her lifespan. Hence, language is represented as an identity determining feature immune to changes in the surrounding cultural context. Previous experiments using the switched at birth task with language as category membership support this conclusion (Hirschfeld & Gelman, 1997).

In addition to social categories, some studies have explored the children's understanding of personality traits (Gelman, Heyman & Legare, 2007; Yuill, 1992, 1998). Thus, it has been shown that personality traits like "mean" and "nice" have rich

inductive potential (Heyman & Gelman, 1999). For example, even 3-year-old children are able to predict that nice people will display a more cooperative behavior than mean people (Heyman & Gelman, 1998), and that two people described like nice or mean will share some preferences regardless their physical appearance (Heyman & Gelman, 2000a). However, children do not seem to believe that personality traits are innate (Heyman & Gelman, 2000b), and when pitted against each other social categories have been shown to have more inductive potential than personality traits for children (Diesendruck & haLevi, 2006).

The children's conceptual status of human kinds regarding natural and artifact kinds has been widely debated. In his original formulation of racial representations, Hirschfeld (1996) proposed that human kinds resemble natural kinds in terms of being objective and intrinsic. A similar formulation was put forward by Gil-White (2001), arguing that ethnic categorizations make use of an innate module for reasoning about biological entities. However, more recent studies call into question the objective nature of children' social representations and overall their similarity to biological categories (Cosmides, Tooby & Kurzban, 2003). Thus, Diesendruck and Eldror (2011), and Diesendruck and Weiss (2015) demonstrated that children represent internal psychological properties but not internal biological properties (e.g. the insides) as definitional of social membership. In other words, people who belong to the same gender, ethnic or racial group are represented as sharing beliefs but not necessarily internal biological mechanisms. Also, Rhodes and Gelman (2009), and Diesendruck, Goldfein-Elbaz, Rhodes, Gelman & Neumark (2013) tested the children's beliefs about the objectivity of different social categories, compared to artifacts and biological kinds by

asking children whether they agree or disagree with alternative categorizations (e.g. a woman being categorize like a man). Overall, they found that the young children's resistance to re-categorize people varies as a function of the type of social category, among other factors. Thus, gender was highly resistant to change, while ethnicity and race were more variable, and new social categories (e.g. shirt-color) were as conventional as artifacts. Therefore, social categories differ from natural kinds in their degree of objectivity. Finally, the social input has been shown to be determinant in the development of social essentialism (Cimpian & Markman, 2011). For example, the use generics facilitates the transmission of social essentialism from parents to children (Rhodes, Leslie & Tworek, 2012), and significant differences in social essentialism has been shown in children across different countries (Diesendruck et al. 2013), cultural contexts (Rhodes & Gelman, 2009a) and racial group membership (Kinzler & Dautel, 2012).

The aforementioned studies suggest that the cultural input has an important role in shaping the representation of social categories. Although this role is now evident the precise connection between cultural input and the construction of human kinds is still an issue of considerable debate. Thus, researchers like Hirschfeld (1996) believe that humans are endowed with an innate capacity to distinguish kinds of people (a folksociology), and language would basically mark what those categories are. On the other side, some researchers (Bigler & Liben, 2010) claim that human kind categories are only the result of cultural experience. In this context, infant studies are crucial to bring insight into this debate. According to Hirschfeld (1996), a pure cultural perspective would be undermined if the emergence of human kinds is traced back to a point as early as the emergence of natural kinds, which are supposed to be less permeable to the social

input. Therefore, finding evidence of an early sensitivity to the social organization and the representation of different kinds of people would support the hypothesis of domain-specific constraints for social essentialism.

Some studies with infants in the first year of life demonstrate an early sensitivity to gender and racial cues in human faces. Thus, 3-month-old infants show a preference to look at female faces and own-race faces (Bar-Haim, Ziv, Lamy & Hodes, 2006; Quinn, Yahr, Kuhn, Slater & Pascalis, 2002), demonstrating that they are able to distinguish males from females, and across different races. Quinn et al (2002) also found evidence of gender-based categorization in 3-4-month-olds, and Anzures, Quinn, Pacalis, Slater and Lee (2010) found that 9-month-old infants are able to categorize Caucasian faces from Asian faces. All these studies have revealed an important impact of cultural context in the infants' ability to discriminate gender and race between and within categories. In trying to get more compelling evidence of an abstract representation of race and gender, Waxman and Grace (2012) tested 7 and 11-month-old infants in categorization tasks combining faces from different racial and gender groups. For instance, infants were presented with different faces from the same racial group (e.g. black), but combining males and females, then two faces from either the same or different racial group were presented in the test trials. The results show that at 7 months of age infants have an abstract representation of gender, but only at 11 month they display an abstract representation of race.

The fact that very early on infants display a preference to look at own-race faces has opened the question about whether or not this bias reflects a deeper "social preference" to interact with people who belong to the same racial group. In addressing

this question, Kinzler and Spelke (2011) presented an event where a black and a white person offered a toy to the participant simultaneously, and then the children's choice was registered. Three age groups were tested, 10-month-old, 2.5-year-old, and 5-year-old children, in order to track developmental changes. The results show that only 5-year-old children display a preference to interact with same-race people. Therefore, the looking preference for own-race faces in infancy may be diagnostic of social familiarity but not of a social preference.

These results contrast with previous experiments carried out with 10-month-old infants, where using the same toy choice task infants display a social preference to interact with people who speak the same participant's language (Kinzler, Dupoux & Spelke, 2007). This language base preference is also apparent in the selective imitation of older infants (Buttermann, Zmyj, Daum & Carpenter, 2013). When 14-month-old infants are shown a video of two people who speak either a native or a foreign language performing actions with an object (e.g. touching a screen with the forehead), they tend to imitate the action only when observe the person who speaks the native language. This effect, however, has been shown to be mediated by the use of videos in the experimental setup (Howard, Henderson, Carrazza & Woodward, 2015).

Overall, these results indicate first, that language for infants seems to be a more relevant social category than race, and second, that early on infants develop a preference to interact and learn from members of the same social group (Dunham, Baron & Banaji, 2008). Kinzler and Spelke (2011) interpret this finding from a nativist perspective, as showing that humans may have evolved the capacity to use language but not race like a valid predictor of group membership or coalitions. This interpretation is supported by

studies with children and adults, revealing that patterns of cooperation and competition are better indicators of social membership than race (Cosmides, et al. 2003), and have rich inductive potential (Rhodes & Brickman, 2011).

Infants have also been shown to be able to distinguish agents based on personality trait information. Namely, between mean (or hinderers) and nice agents (or helpers). In a seminal study, Kuhlmeier, Wynn and Bloom (2003) discovered that 12-month-old infants can predict that an agent will approach to another agent who has been helpful before in accomplish a goal (e.g. reach the top of a hill), while they show surprise if approaches to the hinderer agent. In several additional experiments have been shown that infants also prefer to interact with agents who display a cooperative behavior (Hamlin, Wynn & Bloom, 2007), and with agents who have punished antisocial others (Hamlin, Mahajan, Liberman & Wynn, 2013). It has been demonstrated that these social evaluations are based not on the identification of patterns of behavior but on a mentalistic evaluation (Hamlin, 2013a; Hamlin, 2013b; Hamlin, Ullman, Tenenbaum, Goodman & Baker, 2013). According to Wynn (2008), these results suggest that a "moral sense" is operational early in the first year of life. This system allows to differentiate "good" people from potentially harmful based on the patter of cooperative behavior they display.

Despite being relevant to the discussion about the development of human kinds in infancy, any of the aforementioned studies have tested directly whether or not those social representations are organized like kind concepts. More compelling evidence for this hypothesis come from a study carried out by Powell and Spelke (2013), where 8-month-old infants were shown to be able to infer that members of the same social group may display similar behaviors. Critically, infants were also able to infer that members of

the same social group share the same preferences, an internal psychological property. This finding resembles somewhat the results obtained by Diesendruck and Eldror (2011) with older children about inferences of internal properties across members of the same ethnic group.

CHAPTER 3

THE PRESENT RESEARCH

Even though important progress has been made in revealing how infants categorize their social world (for example, based on race, language or moral dispositions), few developmental studies have explicitly undertaken the project of determining how those social categories are organized in the infants' mind. In the current research, we want to address the category of moral dispositions because this is one of the more studied social categorizations in infancy and there is good evidence that older children tend to essentialize people based on their moral behavior (Heyman & Gelman, 1999; Heyman & Gelman, 2003). Therefore, in this research we address the following question: *Do infants possess a kind concept for an agents' moral disposition?* As has been shown before, representing categories like kinds rather than arbitrary classes allows people to reason about an agents' identity in terms of unobservable properties. Therefore, if it is true that moral dispositions are organized like kind concepts, such organization in the infants' mind should have consequences in the way they reason about the agents' identity across different situations, by rendering the agents' social (moral) membership as more identity-determining than their behavioral or external properties.

In the current research, we want to explore the possibility of an early representation of moral dispositions like kind concepts by testing two specific predictions. First, when information about the agents' moral disposition is available, infants should weigh this information more than the overall agents' appearance to keep track of their *individual identity*. In other words, a change in the type of moral disposition that an agent displays in a social event should be highly diagnostic of a change in the

number of agents participating in the event, regardless similarities they may display in their appearance. Second, when information about the agents' moral disposition is available and the agents' insides are visible, infants should use the insides rather than the external properties to keep track of the agent's *categorical identity*. In other words, internal, "non-obvious" properties should be a more reliable indicator than external properties of what type of social agent is being observed.

Both predictions will be further elaborated in the introduction section of each study. However, it is important to clarify that the plausibility of both predictions derive from three pieces of evidence presented in the previous background research section. First, from around 6 years of age children believe that social categories (including personality traits) are defined by internal rather than external properties (Diesendruck & Eldror, 2011). Second, the pattern of infants' reasoning in categorization, individuation and inductive inference tasks across different conceptual domains (artifacts and natural kinds) demonstrate that early on they expect the world to be populated with 'kinds', meaning that underlying properties define the category membership of some entities (see Csibra & Shamsudheen, 2015). Third, infants are able to distinguish agents with positive social dispositions ("helpers") from agents with negative social dispositions ("hinderers"; Hamlyn, Wynn, & Bloom, 2008). This fact suggests that for infants a moral disposition is a salient property of an agent's behavior, possibly because of the adaptive significance it confers for the establishment of social coalitions (see Rhodes & Brickman, 2011). These three pieces of evidence together suggest that essentialist reasoning is widespread across domains in infancy and childhood. Therefore, it is at least plausible that this same type of bias could underlie infants' representations of others' moral dispositions.

CHAPTER 4

STUDY 1: INDIVIDUATION OF AGENTS BY MORAL DISPOSITIONS

4.1. Experiment 1

4.1.1. Introduction

Infants have been shown to distinguish agents based on the moral dispositions they display (Hamlin, Wynn & Bloom, 2007; Kuhlmeier, Wynn & Bloom, 2003). For instance, when 6-month-old infants witness different characters engaged in either helping or hindering actions, they prefer to interact with the character who displayed a cooperative behavior (Hamlin, 2013a; Hamlin, 2013b). It has been argued that this ability to differentiate "nice" and "mean" agents derives from an innate moral sense, allowing people from very early on in life to establish cooperative bonds with perceived in-group members (Hamlin, 2013; Wynn, 2008).

Even though prior research has indicated that infants are able to represent moral behaviors as salient properties of social agents, no research to our knowledge has explored whether they are represented as moral dispositions that are an intrinsic part of an agent's individual identity.

Previous research in social psychology has demonstrated that people represent some social categories and personality traits as highly diagnostic of individual identity (Haslam, 1998; Heyman & Gelman, 2000). For example, children and adults believe that social categories like "race" and "ethnicity" are more important than "occupation" or "skin color" for representing individual identity insofar as these categories are perceived as immutable and objective (Hirschfeld, 1996). Similarly, some personality traits like

"intelligence" or "talkativeness" are represented as being pervasive and deeply rooted, in contrast to other more transient characteristics like "activeness" or "reservedness" (Haslam, Bastian & Bissett, 2004). Whether or not category membership is perceived as central in the representation of identity seems to depend on the causal structure of each category. When a category, either social or biological, is represented as having unobservable properties that are causally responsible for other visible properties or distinctive features it is more likely that people use the membership to that category as highly identity-determining (Ahn, Taylor, Kato, Marsh & Bloom, 2013; Rehder & Kim, 2009). This type of complex and abstract representation has been referred to by a number of philosophers and psychologist as a *kind concept* (Gelman, 2004; Putnam, 1975).

When in development the representation of kind concepts emerges has been an issue of considerable debate (Rakison, 2003; Mandler, 2004). However, some studies support the hypothesis that kind concepts emerge as early as the end of the first year of life. For example, in the classic study of Xu and Carey (1996) 10- and 12-month-old infants were shown an individuation task where one object (e.g. a ball) emerged from behind a screen, stayed in view for about 5 seconds and then went back to behind the screen; the same procedure was followed by a second categorically different object (e.g. a duck) from the opposite side. The results of this study showed that 12-month-old infants but not 10-month-olds represented two objects behind the screen, as evidenced by their increased looking when witnessing only one object on the stage once the screen was raised. According to Xu and Carey (1996), this result suggests that by 12-months of age infants use category membership (e.g., 'duck') as a more reliable cue of object identity than featural information (e.g., color or shape).

Although both category membership and featural information are strongly correlated, further work has demonstrated that infants use the former as the main cue for individuating objects (Xu, Carey & Quint, 2004). For example, infants who observe the sequential appearance and disappearance of two objects that vary *within* a basic-level kind category (e.g. a sippy cup and a coffee mug) respond as if they only represent a single object behind the screen, even though the two objects could be easily distinguished by the different surface properties they possess. In contrast, infants who observe objects that vary *across* basic-level kinds (e.g. a cup and a ball) represent that there are two objects involved in the event. Other work has replicated this result adding more stringent controls of similarity in the objects' appearance (Kingo & Krojggard, 2011), and extended the findings to 9-month-olds using a reaching paradigm rather than typical looking time measures (Xu & Baker, 2005).

More recent studies in object individuation have demonstrated that kind categories are widely used as a central component in the infants' representation of object identity. For example, one study determined that 10-month-old infants individuate two objects behind a screen if the artifacts are associated with two different functions (Futo, Teglas, Csibra & Gergely, 2010). Similarly, 10-month-old infants individuate two objects if one of them displays self-propelled motion and agentive features (e.g. a worm) and the other looks like a typical inanimate object (e.g. a box; Surian & Caldi, 2010). Together these studies suggest that early on infants tend to disregard visible properties in favor of category membership and non-visible properties (e.g. an artifact's function and agency) when keeping track of object identity over time.

The current research aims to extend the previous findings about the development of kind representations to the domain of social categories by testing whether or not infants are able to individuate agents based on the moral dispositions they display. This investigation will provide insight into whether infants' representations of moral categories are relatively abstract. An early emergence of a kind representation would indicate that for infants, "mean" and "nice" are not just categories with distinctive patterns of behavior and intentions, but the result of intrinsic and unobservable properties common to other agents who exhibit the same type of moral disposition.

Our methodological approach combines the classic object individuation task (e.g., Xu & Carey, 1996) and a recent task designed by Hamlin and Wynn (2011) where a character struggles to open a transparent box. In all of the experiments reported here the subjects observe characters emerging two times from behind a screen to demonstrate same or different sequences of social behavior towards another agent. The main hypothesis is that 11-month-old infants will individuate two agents only when they witness social behaviors with different moral dispositions, regardless of the physical appearance these agents exhibit or other low-level cues. We work with this age range because previous studies have shown that the infants' ability to use kind concepts to individuate objects emerges around 10-12 months of age.

4.1.2. Method

4.1.2.1. Participants

Sixteen 11-month-old infants participated in this experiment (M = 11 months, 3 days, SD = 8 days). Half of them were girls. All infants were recruited from the Amherst,

Massachusetts area. Eleven additional infants participated but were excluded from analysis because of fussiness (2) and experimental error (9). [1]

4.1.2.2. Materials

Infants sat on their parent's lap facing a black stage measuring 118 cm. wide x 75 cm. high (see Appendix A for examples of the stimuli and the experimental setup). The room was dimly lit and parents were instructed to remain silent along the experiment. Infants observed a transparent box (35 cm. wide x 19 cm. deep and 12 cm. high) resting on the center of the stage with two different-colored cubes (5 cm x 5 cm) inside. At the right corner of the stage infants observed a blue screen (25 cm high x 36 cm wide) in a vertical position. There was a gap of 12 cm between the screen and the right frame of the stage and a gap of 17 cm between the screen and the box. Three different puppets were used in the experiment, all measuring 18 x 10 cm. A cow puppet served as the "Protagonist" who struggled to open the box. A pig puppet served as the "Opener" who emerged from behind the screen and helped the Protagonist to open the box by lifting the lid. Another identical pig puppet served as the "Closer" who hindered the Protagonist from opening the box by slamming the lid shut. A black curtain was lowered between trials to hide the stage. Two video cameras recorded events for posterior analyses, one focused on the infant's face and the other focused on the stage.

[1] The complexity of the procedure led to a high number of experimental errors early in our testing. The following is a breakdown of the specific errors: The pig behind the screen was visible for the participant (1), the cow was left on the stage in the test trials (1), the experimenter applied a wrong order of trials (2), the timing of the events was wrong (1), the screen fell in the show revealing the pigs behind (3), or the screen was placed on the wrong position (1).

4.1.2.3. Design and Procedure

Infants were shown 4 baseline trials, 2 familiarization trials, and 4 test trials in a typical violation-of-expectation design as described below.

Baseline Trials. In the Baseline Trials, the curtain was raised revealing an upright blue screen on the stage, then one of the experimenters drew the infant's attention to the stage using infant-directed speech ("Hi [baby's name], look here") before dropping the screen revealing either one or two identical pig puppets. Infants' looking time was recorded and the trial finished when they either looked away for at least two consecutive seconds or after 60 seconds of cumulative looking. This procedure was repeated for a total of 4 baseline trials. The number of revealed objects was counterbalanced across participants (baseline trial block: 1, 2, 2, 1 or 2, 1, 1, 2).

Familiarization Trials. The familiarization trials were modeled from the original box task used in previous demonstrations of infants' moral evaluation (e.g., Hamlin & Wynn, 2011). In the show the Protagonist puppet entered the stage from the left corner and moved to one side of the box. She leaned down to look inside the box three times, then jumped on the front left corner of the box. She then attempted to open the box four times. On the first two attempts she pulled up, lifted the edge of the box a few inches, and dropped it back down. On the third and fourth attempts, she lifted the edge of the lid and lowered it while continuously holding onto the lid, as if the lid was too heavy for her to open. On the fifth attempt, a Pig puppet moved out from behind the left side of the screen, and moved forward next to the box.

During the Opening trial, the Pig puppet jumped on the front right corner of the box, and both the Pig and Protagonist opened the box together. The Protagonist dove down into the box, grabbed one cube, and jumped out of the box to the left side of the stage. The Pig closed the lid and jumped off the box, moving back to behind the screen.

During the Closing trial, the Pig puppet jumped on the frontal right corner of the box, slamming the lid. The Protagonist jumped off the box to the left side of the stage. The Pig puppet jumped off the box, moving back to behind the screen. Both Opening and Closing trials lasted approximately 15 seconds. Once the puppet in the second trial moved behind the screen the Protagonist took the cube she obtained in the Opening trial and ran out of the stage. After 5 seconds the curtain was lowered. Opening and Closing trials were counterbalance across participants.

Test Trials. The first phase of the test trials was identical to the familiarization trials, with infants observing both Opening and Closing trials. In the second phase, once all actions stopped one of the experimenters drew the infant's attention to the screen using infant-directed speech ("Hi [baby's name], look here") and she dropped the screen revealing either one or two identical pig puppets. The Trial Outcome (blocked: 1, 2, 2, 1 or 2, 1, 1, 2) and Trial Order (Opening first or Closing first) were counterbalanced across participants. In all four test trials infants observed the transparent box with one cube inside beside the screen. The duration of the infants' looking time was coded by two independent observers who were blind to the conditions. The inter-observer agreement was high ($r = .96$).

4.1.3. Results and Discussion

Preliminary analyses found no effects of sex, Trial Outcome (1 object or 2 objects first) or Trial Order (Opening first or Closing first); therefore, these variables were collapsed in subsequent analyses. A 2 (outcome: one or two objects) X 2 (trial type: baseline or test) repeated measures analysis of variance (ANOVA) yielded no significant main effect for Outcome, $F(1, 15) = .01$, p = .92, eta = .001, and Trial Type, $F(1, 15) = .31$, p = .59, eta = .02. Importantly, this analysis revealed a significant interaction between Outcome and Trial Type, $F(1, 15) = 13.4$, p = .002, eta = .47, which resulted from longer looking times toward Two Object outcomes (M = 9.81 s., SD = 3.89 s.) than One Object outcomes (M = 7.56 s., SD = 2.93 s.) in the Baseline Trials, and longer looking times toward One Object outcomes (M = 10.5 s., SD = 4.54 s.) than Two Objects outcomes (M = 8.09 s., SD = 2.92 s.) in the Test Trials. Planned comparison t-tests of one- versus two-object outcomes revealed a significant difference in the baseline ($t(15) = -3.2$, p = .006, two-tailed) and a marginally significant difference in the test trials ($t(15) = 185$, p = .08, two-tailed). A total of 12 out of 16 infants had a larger preference for two objects on the Test Trials than on the Baseline Trials (p = .04, via a binomial test).

The results of this experiment suggest that 11-month-old infants succeeded in individuating two agents behind the screen, although both puppets involved in the helping-hindering actions displayed the same external properties. These results are important for two reasons. First, they support the hypothesis that at the end of the first year of life infants represent moral dispositions as highly identity-determining. For infants at this age, observing two different moral actions at different times is more likely to be interpreted as two individuals than only one agent who has changed their moral

disposition towards another. Second, these results add evidence for the relatively early emergence of kind concepts in the first year of life. Previous studies have reported that as early as 10 months of age infants tend to use abstract and non-observable information like more diagnostic of agents' identity than other more accessible properties. Along the same vein, the current study demonstrates that infants are able to use abstract properties like moral dispositions to keep track of the agents' identity. Thus, the representation of kind concepts could be an early achievement in several domains, including social categories.

Even though in the current experiment infants succeeded in individuating agents, this effect may be the result of infants' ability to individuate based on the number of actions they observe rather than being based on different moral dispositions. Previous studies have reported that 6-month-olds are able to individuate and enumerate actions from continuous motion (Sharon & Wynn, 1998; Wynn, 1996). For example, when infants observe a sequence of 2 identical actions (jumps) they dishabituate when observing 3 actions, even if both sequences have the same duration. Therefore, an alternative explanation for the pattern of results reported here is that infants count 2 actions along each trial (one for the Opener and another for the Closer) and then they expect a correspondence between the number of actions and the number of puppets behind the screen, resulting in longer looking times for 1 object than for 2 objects outcome in the test trials. To test for this possibility a second experiment was run using the same box task but presenting 2 identical moral dispositions in each trial, either helping or hindering actions. If in the first experiment infants individuate actions based only on numerical information, the pattern of results should be replicated in the second experiment.

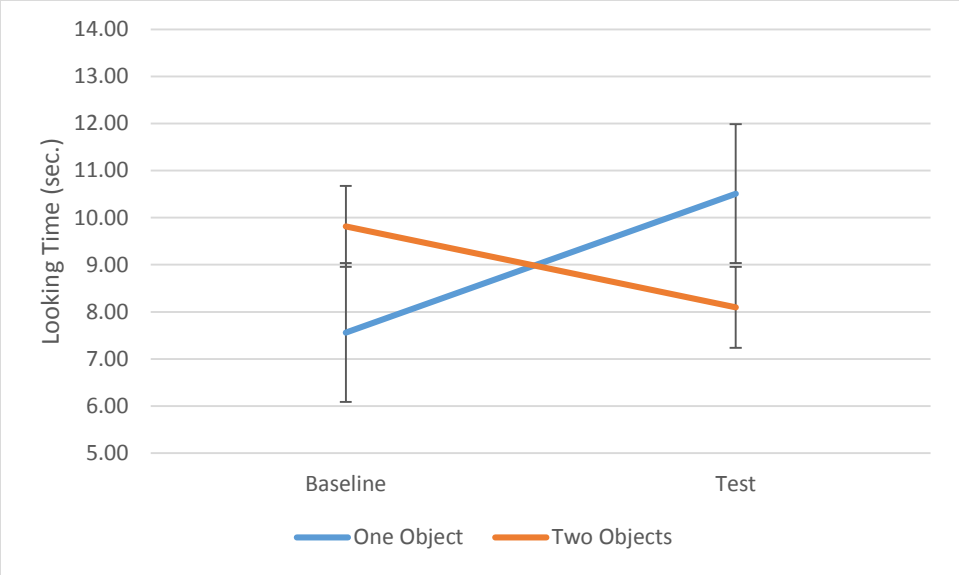


Figure 1. Mean Looking-Time Results Experiment 1.

4.2. Experiment 2

4.2.1. Method

4.2.1.1. Participants

Sixteen 11-month-old infants participated in this experiment (M = 11 months, 2 days, SD = 7 days). Half of them were girls. All infants were recruited from the Amherst, Massachusetts area. Three additional infants participated but were excluded from analysis because of fussiness (2) and experimental error (1).

4.2.1.2. Materials, Design, and Procedure

The materials, design and procedure for the second experiment were the same for that of Experiment 1, except that both social actions infants witnessed were identical in the pattern of motion and in the moral disposition they display (both helping actions or

both hindering actions). The type of moral disposition infants observed was counterbalance across participants. In order to be consistent regarding the number of cubes that infants observe in the box across test trials and across experiments, the hindering event started off with only one cube inside the box, and the helping event started off with three cubes inside the box. The result of two helping actions and two hindering actions was always one cube inside the box. The inter-observer agreement of this experiment was high (r = .95).

4.2.2. Results and Discussion

Preliminary analyses found no effects of sex, Trial Outcome (1 object or 2 objects first) or Trial Order (Opening first or Closing first); therefore, these variables were collapsed in subsequent analyses. A 2 (outcome: 1 or 2 objects) X 2 (trial type: baseline or test) repeated measures analysis of variance (ANOVA) yielded no significant main effect for Outcome, $F(1, 15) = 2.08$, p = .17, eta = .012, and Trial Type, $F(1, 15) = .71$, p = .41, eta = .04. This analysis did not reveal a significant interaction between Outcome and Trial Type, $F(1, 15) = .105$, p = .75, eta = .007. Infants spent the same time looking at the 1 and 2 objects outcome in both the baseline trials (M = 8.48, SD = 5.09; M = 9.1, SD = 3.72, for one object and two objects respectively, $t(15) = -.69$, p = .5, two-tailed) and the test trials (M = 7.3, SD = 3.27; M = 8.4, SD = 3.44, for 1 object and 2 objects respectively, $t(15) = -1.1$, p = .28, two-tailed). Finally, a 2 (outcome: 1 or 2 objects) X 2 (trial type: baseline or test) X 2 (Experiment Type: Experiment 1 or Experiment 2) analysis of variance (ANOVA) yielded a significant three-way interaction among Outcome, Trial Type and Experiment Type, $F(1, 30) = 7.02$, p = .01, eta = .19. This

interaction reveals that the pattern of results in Experiment 2 is significantly different from that of Experiment 1.

The results of Experiment 2 show that infants fail to individuate 2 objects behind the screen. Although infants always observe 2 emergences and two separate actions within each trial they do not seem to use this information to infer the number of objects present behind the screen. Previous studies have shown that infants are able to individuate and count actions (Wynn, 1996), however the results of the current experiment show that the number of actions they observe in each trial is not salient enough to represent different agents in a spatiotemporal ambiguous situation. Additionally, in this experiment infants observed conflicting evidence to individuate objects. On the one hand, numerical information indicated two objects behind the screen, and on the other hand featural and social information suggested only one object. This conflict may have increased the uncertainty about the number of puppets behind the screen.

Even though the Experiment 2 rules out the option of object individuation based on the number of actions perceived, there are other two low-level explanations that may account for the results in Experiment 1. First, helping and hindering actions differ not only in the moral disposition they represent, but also in the pattern of motion that those actions display. Hindering actions are characterized by pushing the lid down and helping actions by lifting the lid. Second, helping and hindering actions differ also in the type of first order goal that mediates the social interaction. Namely, a hindering action in the box task requires the intention to close the box, resulting in the representation of that agent as a “closer”, while a helping action requires the intention to open the box, resulting in the

representation of that agent as an "opener". Either of these alternatives, or both together, may be driving the effect observed in Experiment 1 without any commitment with the social interaction among the different characters. In order to test for these possibilities a third experiment was conducted presenting a puppet show with one character opening and closing a box at different times. The Protagonist was eliminated from the show to avoid any interpretation of the events in terms of social interactions. If infants individuate agents based on differences in motion cues and first-order goals, the pattern of results of the first experiment should be replicated in the current one.

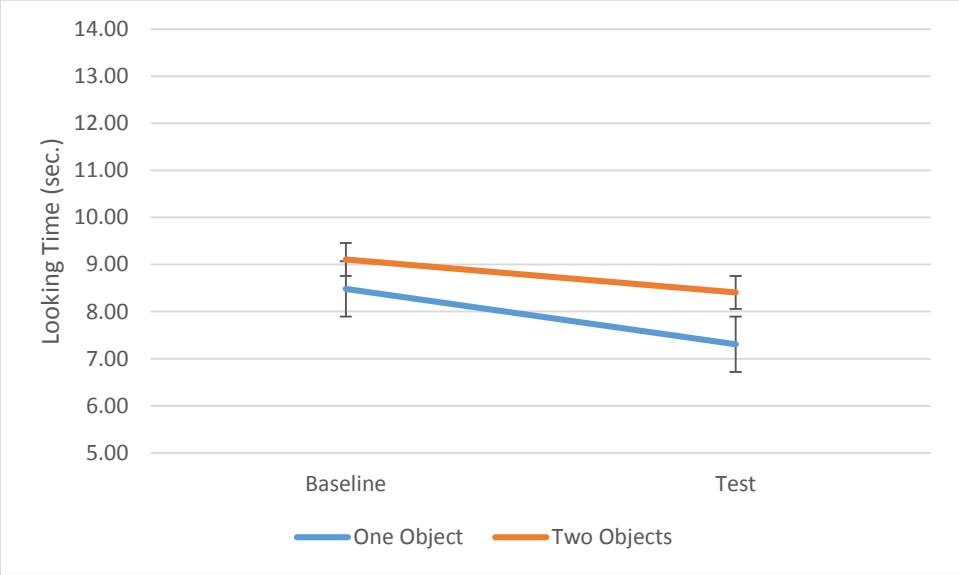


Figure 2. Mean Looking-Time Results Experiment 2.

4.3. Experiment 3

4.3.1. Method

4.3.1.1. Participants

Data collection is still ongoing. So far participants are 14 infants out of 16 (M = 11 months, 1 day, SD = 7 days), seven males and seven females. All infants were

recruited from the Amherst, Massachusetts area. One additional infant participated but he was excluded from analysis because of fussiness.

4.3.1.1. Materials, Design, and Procedure

The materials and design of the third experiment will be the same for that of Experiment 1, except that in the Familiarization and Test Trials the Protagonist (the cow) and the cubes inside the box will be removed from the show. The pattern of motion of both the Opening and the Closing actions will be very similar to the pattern of motion used in the previous two experiments. During Opening trials, the Pig puppet will jump on the frontal right corner of the box, pulling up the lid completely backwards. During Closing events, the Pig puppet will grab the lid to close the box in a forward movement. A pause of about 5 seconds between both actions will be used.

4.3.2. Results and Discussion

Preliminary analyses found no effects of sex, Trial Outcome (1 object or 2 objects first) or Trial Order (Opening first or Closing first); therefore, these variables were collapsed in subsequent analyses. A 2 (outcome: 1 or 2 objects) X 2 (trial type: baseline or test) repeated measures analysis of variance (ANOVA) yielded no significant main effect for Outcome, $F(1, 13) = .12$, $p = .73$, eta < .01, and Trial Type, $F(1, 13) = 3.13$, $p = .1$, eta = .22, As in the previous experiment, this analysis did not reveal a significant interaction between Outcome and Trial Type, $F(1, 13) = .038$, $p = .85$, eta < .01. Infants spent the same time looking at the 1 and 2 objects outcome in both the baseline trials (M = 9.74, SD = 5.75; M = 9.89, SD = 5.1, for one object and two objects respectively, $t(14)$

= -.007, p = .99, two-tailed) and the test trials (M = 11.93, SD = 4.93; M = 12.57, SD = 4.32, for 1 object and 2 objects respectively, $t(14) = -.5$, p = .62, two-tailed). Finally, a 2 (outcome: 1 or 2 objects) X 2 (trial type: baseline or test) X 2 (Experiment Type: Experiment 1 or Experiment 3) analysis of variance (ANOVA) yielded a significant three-way interaction among Outcome, Trial Type and Experiment Type, $F(1, 28) = 4.1$, p = .044, eta = .1. This interaction reveals that the pattern of results in Experiment 2 is significantly different from that of Experiment 1.

The partial results of Experiment 3 show that infants fail to individuate 2 objects behind the screen. Although the pattern of motion for Closing and Opening trials are perceptually similar to the pattern of motion of Helping and Hindering events of Experiment 1 infants do not seem to use this information to infer the number of objects involved in the show in the current experiment. These results are in line with previous findings in individuation studies. For example, in the second experiment of Surian and Caldi (2010) 10-month-old infants observed two different animals emerging from different sides of the screen in a typical individuation paradigm. Crucially for the current experiment, both animals displayed different locomotion (e.g. walking, crawling, jumping, and flying). In spite of clear differences in appearance and pattern of motion infants failed to individuate two agents.

The current experiment also rules out the possibility of object individuation based on different first-order intentions; namely, the intention to close and the intention to open a box. Therefore, at least in this particular scenario, 11-month-old infants know that two different non-social intentions are not sufficient evidence to represent two different individuals.

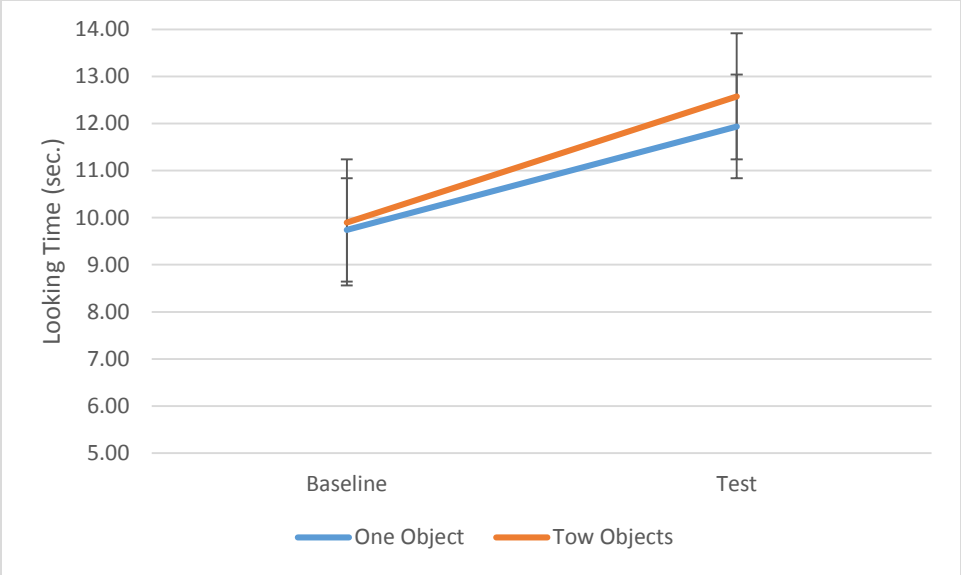


Figure 3. Mean Looking-Time Results Experiment 3

4.4. General Discussion

The current study used an individuation task to investigate whether 11-month-old infants use moral dispositions to keep track of the agents' individual identity. Experiment 1 found that when infants observe two different socio-moral actions, helping-hindering, they individuate two agents, regardless similarities in external properties those agents display. By contrast, in Experiment 2 we found that when infants observe two identical socio-moral actions, either helping-helping or hindering-hindering, they fail to individuate two agents, indicating that infants do not use the perceived number of actions to infer the number of agents involved in the show. Likewise, so far in Experiment 3 infants fail to individuate two agents based on differences in motion and first-order intentions, close and open a box, that resemble the actions infants observed in Experiment 1.

The results of all three experiments together suggest that around the end of the first year of life infants represent moral dispositions as more identity-determining than agents' overall appearance, first-order goals and differences in motion. This may indicate an early bias to represent moral dispositions in terms of different categories, so that antisocial and prosocial behaviors are more readily attributed to different agents than to the same agent. This bias could be associated to the categorical representation of in-group versus out-group members based on the perception of coalitional alliances. Identifying and keeping track of a person as an antisocial individual that is different from cooperative members in a community could be relevant for increasing the likelihood of survival.

The categorical representation of prosocial and antisocial agents seems to lead infants in the current set of experiments to treat moral dispositions as an attribute resistant to change. This is a typical feature of essentialized categories like natural kinds and other social kinds. For instance, people believe that a dog cannot change its identity even if his appearance and behavior are cat-like. This suggests that moral dispositions are represented as kind categories, and therefore possessing a non-observable property responsible for the agents' individual identity over time. Thus, the current study is in agreement with previous kind-based individuation studies of natural and artifact categories, adding evidence for an early emergence of kind concepts.

CHAPTER 5

STUDY 2: INSIDES AND MORAL DISPOSITIONS

5.1. Experiment 1

5.1.1. Introduction

A growing body of literature suggest that by 5 years of age children outweigh internal properties over external properties when reasoning about social categories. For example, children this age believe that people who belong to the same social group may share both internal biological properties and internal psychological characteristics (Diesendruck & Eldror, 2011), and they believe that internal properties are more important for inferring peoples' social membership than labels and other external properties (Diesendruck & Weiss, 2015). This evidence has led some researches to suggest that children represent peoples' internal properties as a proxy for an essential attribute that is responsible for their category membership and social identity (Diesendruck & Eldror, 2011).

What are the developmental origins of this abstract representation? Recent investigations suggest that language, and in particular the use of generics in pedagogical contexts, is an important factor in the transmission of social essentialism in preschool-aged children (Diesendruck & Deblinger-Tangi, 2014; Rhodes, Leslie & Tworek, 2012). However, it is unknown how children reason about the role of internal properties in social categories before language acquisition. Some researchers have suggested that domain-specific cognitive biases may operate early in infancy to shape the children's representation of social categories. For example, according to Gil-White (2001) social essentialism is the by-product of reasoning about people in terms of biological kinds. In

particular, he proposes that due to surface similarities people represent ethnic groups like different "species." Because animals have been shown to be highly essentialized across different cultures (Gelman, 2003; Keil, 1989), ethnic groups turn out to be conceptualized in a similar fashion. Likewise, Hirschfeld (1996) proposes that humans are endowed with an innate capacity to distinguish kind of people (a "folksociology") and language would basically mark what those categories are. From either perspective – an innate social module or analogical transfer—the emergence of social essentialism could be traced back in development well before 4-5 years of age when most of the studies have reported the presence of essentialist beliefs in children.

Some recent studies with infants have shown an early appreciation of the importance of internal properties when reasoning about biological entities. For example, Welder and Graham (2006) discovered that 14-month-olds are more willing to categorize objects based on internal properties when they look like animate objects (e.g. with eyes) but not when they look like inanimate containers. Newman, Herrmann, Wynn, and Keil (2008) also demonstrated that when 14-month-old infants are presented with self-propelled semitransparent objects displaying different types of motion they tend to associate the color of internal properties with a particular type of motion. Crucially, in a second experiment these authors demonstrated that when objects lack of self-propelled motion infants do not show a significantly higher preference for internal over external properties. The importance of self-propelled motion in the infants' representation of internal properties has been further investigated in work showing that 8-month-olds infer that self-propelled and agentive entities should have something inside (Setoh, Wu, Baillargeon & Gelman, 2013). Importantly, when objects lack either self-propelled

motion or agentive features infants do not show any expectation about objects' internal properties. The authors interpret these findings as showing that infants believe that internal properties are causally responsible for both animals' agentive features and self-propelled motion. Finally, using semitransparent objects Taborda-Osorio and Cheries (2015) show that 13-month-old infants are able to individuate agents based on the color of their internal features, while they disregard color properties in the agents' external appearance, suggesting that infants this age represent the agents' internal properties as more diagnostic of individual identity than external properties.

Although these previous studies show an early bias toward internal properties, none of them test whether internal properties play a role in the infants' conceptualization of social categories. The main hypothesis of the current study is that if infants essentialize social categories like older children do, they may be biased to outweigh internal properties over external properties to keep track of the agents' social membership. This pattern of reasoning would be in agreement with how infants have been shown to categorize objects in other domains. For instance, Ware and Booth (2010) demonstrated that 17-month-olds categorize artifacts based on the perceptual properties they display only when these properties are diagnostic of deeper functional properties. Therefore, infants in the second year of life seem to have the notion that non-obvious properties are more reliable cues of kind membership than obvious external features, like shape or color. The current study aims to bring some insight into how this same pattern of reasoning may operate in the domain of social kinds at the onset of language acquisition.

Typically, developmental studies in social kinds with preschool-aged children are focused on categories such as race and ethnicity (Rhodes & Gelman, 2009a). However,

these have been shown to vary significantly across different cultures and depend in to a great extent on the social input that the child receives (Diesendruck & Goldfein-Elbaz, 2013; Kinzler & Dautel 2012; Diesendruck & Deblinger-Tangi, 2014). By contrast, infants as early as 6 months of age have been shown to distinguish agents based on their socio-moral dispositions. In concrete terms, infants show a preference for agents who display cooperative behavior over agents who display antisocial behavior (Hamlin & Wynn, 2011; Hamlin, Wynn, & Bloom, 2008; Wynn, 2007), and they expect other agents to show the same preference (Kuhlmeier, Wynn & Bloom, 2003). Therefore, if infants have an essentialist representation of these moral categories they may be willing to use visible internal properties as a more reliable cue of kind membership than external features. To test this hypothesis, a replica of the classic "hill task" will be used with 14-month-olds. We use this age range because previous studies have shown that infants display a bias toward internal properties at around 13-14 months of age. In our "hill task" infants observe one character trying to reach the top of a hill unsuccessfully, then two other agents with different geometric shapes and color are shown either helping or hindering the main character to fulfill its goal. In the current study, the geometric figures will be replaced with semitransparent objects with internal and external properties of the same color. During test trials, two different characters will be presented in front of the infant. One will display the internal properties of the helper character and the external properties of the hinderer, while the other will display the opposite combination. If infants use internal properties to keep track of the agents' membership, they may prefer to interact with the character with the same internal features of the original helper agent, even though it displays different external features.

5.1.2. Method

5.1.2.1. Participants

Sixteen 14-month-old infants participated in this experiment (M = 14 months, 14 days, SD = 8 days). Half of them were girls. All infants were recruited from the Amherst, Massachusetts area. Six additional infants participated but were excluded from analysis because of fussiness (2) and failure to choose (4).

5.1.2.2. Apparatus

Infants sat on their parent's lap facing a display stage of 120 cm. wide x 95 cm. high. (see Appendix B for examples of the stimuli and the experimental setup). The room was dimly lit and parents were instructed to remain silent along the experiment. The display stage had a white background made of foam and a green base made of wood, with a 4 inches canal rising from lower left to upper right corner, resembling a hill. It had a small plateau one-third of the way up and a second at the top. The climber character was a blue circle made of wood with googly eyes placed on the upper half looking toward the top of the hill. The other four characters were plastic transparent cans 20 cm. high and 10 cm. wide. Each character had two googly eyes attached on the upper half of the can. Two identical pyramidal structures made of balls of cotton were placed one in the bottom of the can (the internal property) and the other on the very top, attached to the lid (the external property). A white paper sheet was folded inside the can, covering the back and the upper half of the can. Two of these characters (the helper and the hinderer) had the same color properties inside and outside: one with red cotton and the other with yellow

cotton. The other two characters (the test characters) had a contrasting combination of color properties: one with yellow cotton inside and red cotton outside, and the other with red cotton inside and yellow cotton outside. A white foam sheet of 50 cm long and 20 cm wide was used to place the two test characters in the test trial. A black curtain was lowered between trials to hide the stage. Two video cameras recorded events for posterior analyses, one focused on the infant's face and the other focused on the stage.

5.1.2.3. Procedure

The procedure of the current experiment was modeled from the original Hamlin, Wynn and Bloom (2008) study. The curtain was first raised and lowered three times without any character in the display stage. In each familiarization trial the climber character wiggled for one second while on the left bottom of the stage, then climbed to the middle plateau where paused and wiggled again for one second. The climber subsequently attempted twice to reach the top of the hill, each time falling back to the middle plateau. On a third attempt, the climber was either pushed up to the top by the helper, or pushed down to the bottom by the hinderer. In the helping event, the helper entered to the display stage from the lower left, moved up the incline and pushed the climber twice, each time pushing it closer to the top until the climber reached the upper plateau. Once on the top the climber wiggled while the helper went downhill to the bottom plateau and paused. In the hinderer event, the hinderer entered to the display stage from the upper right, moved down the incline and pushed the climber twice, each time pushing it closer to the middle plateau. The climber then moved downhill to the lower plateau, and the hinderer moved back to the top of the hill and paused. Total duration of

each event was 10 sec. Infants were exposed to three hindering events and three helping events.

In the test trial, the experimenter presented both test characters 40 cm. apart on a board, and asked "Can you show me who is the nice one?" Then she moved the board forward and looked down. Infant's choice was defined as the character touched first, as judged by the (blind) coder, with the constraint that the infant had to be looking at the toy during or immediately preceding the touch. The color of the hinderer and helper characters, the order of habituation trials, and the left-right position of the test characters were counterbalanced across participants.

5.1.3. Results and Discussion

Preliminary analysis did not reveal order effects of the position of the test characters, habituation trials or the order of color presentation. Results show that infants robustly chose the character with the same color inside as the helper character in the familiarization trials (13 out of 16, $p = .02$, two-tailed, by a binomial probability test). This result, first, replicates previous findings where infants this age and younger choose the helper character after being exposed to socio-moral events with a helper and a hinderer character. Second, the current study extends previous findings by showing that infants are able to use the color of internal properties to keep track of agents' socio-moral membership. In other words, infants identify the "nice" character based on the internal properties while they disregard the external properties the character displays. Thereby, this study provides support for the hypothesis that moral categories are represented like intrinsic and essentialized categories at the onset of language acquisition.

Why do infants use internal physical properties as a more reliable cue of moral disposition than equally visible external properties? One possibility is that infants may be biased to represent moral categories, and other categories in diverse conceptual domains, as kind categories. That is, infants reason about social agents under the assumption that non-obvious properties are causally responsible for the pattern of behavior and moral dispositions they display. In this regard, the possession of some external characteristic features (such as skin color for race, or patterns of behavior for moral categories) is not the reason by which an entity belongs to a category, but rather the effect of some deeper, typically unobservable, causal essence. Just as agents' internal properties have been shown to be represented from very early on as an important biological property (Setoh, et al., 2012) and more relevant for agents' identity than external properties (Taborda-Osorio & Cheries, 2015), infants in the current study may use those internal properties as a proxy for an essential moral disposition. It remains to be seen how early this reasoning toward sociomoral dispositions emerges, and what is the developmental trajectory along childhood. Although infants as early as 6 months of age seem to distinguish "mean" from "nice" agents it is unclear if they would be equally willing to associate those dispositions with internal properties. For this infants would have to assume first that animate agents have insides, and second that sociomoral dispositions are causally motivated by intrinsic properties. The earliest evidence of attribution of internal properties to animals is at 8 months of age, so it is feasible that even before the first year of life infants exhibit a similar bias toward internal properties as 14-month-olds do.

A central piece in the previous interpretation is that infants pay more attention toward agents' internal properties because they have a more relevant causal role in the

representation of moral categories. As it has been demonstrated in several experiments about categorization, adults and children tend to categorize objects based on features that have a causal role in supporting the presence of other features (Ahn, 1998; Ahn, Gelman, Amsterlaw, Hohenstein, & Kalish, 2000; Rehder, 2003). Therefore, a more direct way to test the hypothesis that 14-month-olds represent internal properties as a proxy for an essence would be to determine whether or not infants attribute causal potency to the internal properties they perceive in the characters involved in socio-moral behavior. To test this hypothesis a second experiment will be run where infants, prior to the habituation, witness one of two types of familiarization trials: either an event of the insides being removed or an event of the outsides being removed. In both events, once the property has been removed infants will observe the character moving up and down along the hill on the display stage. The goal with this manipulation is to demonstrate to the infant that either the internal or the external property is not causally relevant for the pattern of motion the agents display. In the case of removing the insides we expect that by weakening the role of this internal property in infants' representation of biological agency and individual identity, they will be less biased to use internal properties like the main cue of kind membership. By contrast, in the case of removing the outsides infants should still be willing to use the internal properties like the main cue of kind membership. Experiment 2 also addresses an alternative explanation for the pattern of results in Experiment 1. Infants may prioritize the internal features because they are placed on the bottom part of the toy where animals have mobile parts (e.g. the month and legs). If this is true, we should replicate the results in both conditions (insides removed and outsides

removed) because in both cases the internal part was visible on the bottom along the helping and the hindering trials.

5.2. Experiment 2

5.2.1. Method

5.2.1.1. Participants

Data collection is still ongoing. So far participants are 12 infants (M = 14 months, 11 days, SD = 6 days), six males and seven females. Infants were randomly assigned to either Insides Removed condition (6) or Outsides Removed condition (6). All infants were recruited from the Amherst, Massachusetts area. One additional infant participated but he was excluded from analysis because of fussiness.

5.2.1.2. Apparatus

The display stage, the climber character, and the test characters were the same as in Experiment 1. The helper and the hinderer character had the same overall appearance but the bottom of the can has a hole through which the internal material can be removed, and the external material is attached to the top with small flat magnets to facilitate its removal.

5.2.1.3. Procedure

Prior to the habituation trials, separate groups of infants witnessed two familiarization trials (one for the helper character and the other for the hinderer character), presenting either an event of temporarily removing the insides (the Insides

Removed Condition) or an event of temporarily removing the outsides (the Outsides Removed Condition). In both events the trial started with the character placed on the middle plateau, a hand wearing a white glove showed up through the canal while the experimenter called the infant's attention with infant-directed speech ("look here [baby's name]"). Next the experimenter proceeded to remove either the internal or the external material, pausing for about one second while moving the now detached property 10 cm apart from the character, after which either the external or internal material was withdrawn from the display stage through an opening by the edge of the stage. With the property now absent, the character was shown climbing the hill all the way up, then moving all the way down to the bottom, and then climbing back to the middle plateau where it pauses. This sequence of events was then repeated for the other-colored puppet. Then, either the internal or the external properties were put back in the character out of the infant's view. All other six habituation trials and the test trial occurred in the same fashion as in Experiment 1.

5.2.2. Results and Discussion

Expected results. We predict that infants will not show a reaching preference towards either test trial object in the Insides Removed Condition. In contrast, we expect to replicate the significant reaching preference we observed in Experiment 1 in the Outsides Removed Condition, in such a way that infants choose the helper character based on the color of its internal properties. These results would indicate that only when internal properties are represented like functional biological properties with causal

potency infants are willing to use those properties to keep track of the agent's socio-moral membership.

Current results. In the Insides Removed Condition 3 out of 6 infants chose the character with the same color inside as the helper character in the familiarization trials, while in the Outsides Removed Condition 4 out of 6 infants chose the character with the same color insides as the helper. Although overall this pattern of results is in agreement with what it was predicted, 6 participants in each condition is still a too small sample size to conclude anything. These results could also indicate that infants chose randomly in both conditions, suggesting that the experimental manipulation disrupted the identification of the helper character in both conditions. More subjects will be ran to tease apart both possibilities. However, so far we do not have evidence to support the hypothesis that infants attribute causal potency to the agents' internal properties.

CHAPTER 6

GENERAL DISCUSSION

The partial results of Study 1 suggest that 11-month-old infants are able to individuate agents based on sociomoral information. In Experiment 1 infants expected two individuals behind the screen when they observed two different sociomoral behaviors at different times. Experiment 2 ruled out the alternative low-level explanation of individuation based only on numerical information. So far, Experiment 3 rules out the possibility of individuation based on differences in motion or differences in first-order goals. If the pattern of results in Experiment 3 follow our prediction and infants fail to individuate objects, this would support the hypothesis that the representation of agents as either prosocial or antisocial is more identity-determining than similarities in featural information.

The partial results of Study 2 suggest that 14-month-old infants are able to use agents' internal properties as a more reliable cue to distinguish prosocial from antisocial agents. Experiment 1 showed that infants identify the prosocial character based on the color of its internal features. Experiment 2 will determine if this bias is the result of a causal understanding of internal properties associated with the generation of agentive behavior. If the pattern of results in Experiment 2 follow our prediction and the identification of the prosocial agent is disrupted in the "insides removal condition" and preserved in the "outsides removal condition", this would support the hypothesis that internal properties are more essential than external properties by virtue of their causal role in the generation of prosocial and antisocial behavior.

As a reminder, the main hypothesis of the current set of studies is that at the end of the first year of life infants possess kind concepts for representing sociomoral dispositions. That is to say that infants, first, represent "mean" and "good" traits as general and abstract categories composed by indefinitely many instances (Prasada, 2012). Thus, examples of prosocial and antisocial behavior could be interpreted as qualitatively different from each other in a fundamental way. Second, whether an agent displays either a prosocial or an antisocial behavior depends mainly on the possession of an internal non-visible property that is causally responsible for those behaviors. In other words, being "mean" or being "good" does not depend on contingent external properties of a particular person, but on internal attributes that "good" and "mean" people share. Previous studies have shown that children use such an abstract and complex representation to reason about personality traits and, in particular, about sociomoral traits. For example, children find "nice" and "mean" traits more inductively powerful than the external appearance of people (Heyman & Gelman, 2000). They infer that people with the same personality trait share similar emotions, behaviors and thoughts regardless their external appearance. This suggests that for children the labels "nice" and "mean" indicate deep similarities. Additionally, children believe that positive personality traits tend to remain more stable across time and different situations than neutral attributes (Diesendruck & Lindenbaum, 2009). Therefore, if the origins of such an abstract kind-based representation can be traced back to the first two years of life, infants should represent "mean" and "good" dispositions as relatively stable over time, qualitatively different each other, and they may base this reasoning on the attribution of internal properties.

Overall, the current set of studies support our main hypothesis because moral dispositions are represented as a central component of individual identity (Study 1), and because the distinction between prosocial and antisocial agents is mainly based on internal properties (Study 2). However, these findings allow two different interpretations. First, similar to natural kind concepts (e.g. animals) infants may believe that the moral categories "good" or "mean" are objective, mutually exclusive, and intrinsic kinds. In other words, they believe that there are two types of completely different people in the world, "good" or "mean". This representation would be objective because the category membership of each agent should be discovered through observation. They are mutually exclusive because no agent can be good and mean at the same time. And they are intrinsic because people cannot change their moral category membership over time. Study 1 seems to support this interpretation insofar as infants tend to infer two different characters rather than one when observing different sociomoral actions, so they seem to be represented as mutually exclusive categories. Study 2 also supports this interpretation because internal properties are usually linked to objective kinds for both animals and plants, so that internal properties are stable and do not change over time. This type of "natural kind" representation of moral categories could be the result of early biases to identify uncooperative people as members of external groups, while cooperative people could be identified as members of the same observer's group. Therefore, representing "good" and "mean" as objective and intrinsic kinds associated with in-group and out-group members could be useful for predicting future behaviors (e.g. a mean character will always display an uncooperative behavior in any type of situation).

A second interpretation of the current set of studies is that infants believe that the moral categories "good" and "mean" are extrinsic and graded kinds (Estes, 2003). This type of representation would put moral categories somewhat closer to the children's representation of artifact kinds (Rhodes & Gelman, 2009a) and some social categories such as race and ethnicity (Rhodes & Gelman, 2009b). The representation of "good" and "mean" would be extrinsic because category membership depends on more contextual factors that can change over time (e.g. the social context), and it could be graded because people may be "mean" and "good" at the same time to some extent. It has been shown that infants can take into account contextual factors when evaluating sociomoral actions. For example, although infants prefer to interact with prosocial over antisocial agents, they also prefer antisocial agents who harm dissimilar others (Hamlin, et. al, 2013). Thus, they know that being "mean" or "nice" depends on the previous history of the characters involved. However, as artifacts and some social categories "good" and "mean" could still be considered kind-based representations for at least two reasons. First, because non-observable properties are more identity-determining than observable properties (e.g. the external appearance or patterns of motion). Second, because being "mean" or "nice" could be properties of kinds of individuals, meaning that they are represented as properties of an unlimited group of people (Prasada, 2012). Study 1 supports this interpretation because those experiments suggest that the possession of specific moral dispositions rather than external similarities or perceptual differences in motion drive the infants' numerical expectations. In the same line of reasoning, Study 2 also supports this interpretation because internal non-obvious properties are more reliable indicators of moral category membership than external properties. The main difference between the

first and the second interpretation of these findings is that infants may represent moral kinds as relative rather than absolute categories.

Which of these options is better supported by the developmental literature? Some studies about the understanding of personality traits indicate that adults represent some personality traits (e.g. shy, cold, talkative) as discrete categories, immutable, and biologically rooted (Haslam, et al., 2004). For example, in the case of "shyness" adults believe that people "either have this characteristic or they don't and "it is not easy to change". On the other hand, studies with preschool-aged children suggest that around 4 years of age personality traits like "mean" and "nice" have inductive potential (Heyman & Gelman, 2000). However, they are inductively less powerful than social categories (e.g. ethnicity), children do not believe that they are biologically rooted, and essentialist beliefs about personality traits are less coherent in young that in older children. Accordingly, some researchers have suggested that the representation of personality traits like essentialized natural kinds (e.g. objective and intrinsic) could emerge in adulthood as a result of exposure to biological theories in the school. Therefore, younger children and infants may have a less essentilized representation of personality traits, including "mean" and "nice".

Another reason by which we could be skeptical about a strong interpretation of the current set of findings in terms of natural kinds is because the people's moral dispositions tend to vary as a function of the social context. For example, being "mean" or being "nice" are dispositions relative to the recipient of those actions (e.g. a person could be "mean" and "nice" at the same time with different people). By contrast, the category membership of an animal does not change as a function of the context.

Therefore, if infants represent moral dispositions as extrinsic kinds, they may allow for some flexibility in the attribution of different dispositions to different individuals, but they may still be identity-relevant properties (see Pomiechowska, Tatone, & Csibra, 2016).

Regardless of whether infants represent sociomoral categories as intrinsic or extrinsic kinds the current set of studies provide two new insights about the origins of sociomoral reasoning. First, around the start of the second year of life infants believe that individuals display a coherent set of sociomoral behaviors in a particular context (e.g. toward the same character). When infants detect two opposite sociomoral behaviors in a featural and spatiotemporally ambiguous situation they are biased to represent two different individuals. Second, the heavier weighing of internal over external properties of social agents in Study 2 may signal the beginnings of essentialist reasoning of sociomoral dispositions and social categories observed in older children and adults. Despite differences in external appearance infants may believe that something intrinsic is more determinant of the type of agent that they observe.

Several other issues deserve further investigation in the future. First, to what extent do infants' representations of sociomoral categories have inductive potential? A characteristic feature of essentialized kinds is that knowing the category membership of an object allows people to infer new properties of that object. For instance, knowing that an animal is a mammal allow people to infer that is "warm-blooded" and produce milk. It has been shown that 4-year-olds believe that "mean" and "nice" categories have more inductive potential than the person's appearance. If infants have a similar intuition they may be able to infer that "nice people" tend to share behaviors and preferences among

them. They also may infer that "nice" or "mean" people will be relatively consistent in their intentions and behaviors across different situations. For example, if infants believe that a prosocial individual has an underlying positive motivation to help others, they may predict a wide range of prosocial behaviors in diverse situations. A second research question has to do with differences in the conceptualization of prosocial and antisocial behavior. In particular, are both sociomoral categories, "good" and "mean", equally essentialized in infancy? Some previous studies with older children suggest that positive personality traits could be more strongly essentialized than antisocial behavior (Diesendruck & Lindenbaum, 2009). Thus, children believe that positive traits (e.g. being sociable) are more stable over time and across situations than negative traits (e.g. being a loner), which could indicate that positive traits are more essentialized than negative traits. Therefore, positive traits could be conceptualized as an essential part of "human nature". Future research should clarify if this intuition emerges even earlier in infancy. A third open question has to do with whether infants can individuate agents based on other social dispositions besides sociomoral behavior. Agents are engaged in multiple types of social interactions and some of them are salient for infants early on. For example, infants are able to distinguish chasers from chasees (Rochat, Striano, & Morgan, 2004) and dominants from subordinates (Thomsen, Frankenhuis, Ingold-Smith & Carey, 2011). The extent to which infants represent different types of social roles as identity markers has not been previously explored.

Finally, although in psychology the notion of essentialism has been traditionally connected to the distinction of kinds from non-kinds, it is still unclear the extent to which children and infants represent sociomoral behaviors as a property of a general kind, rather

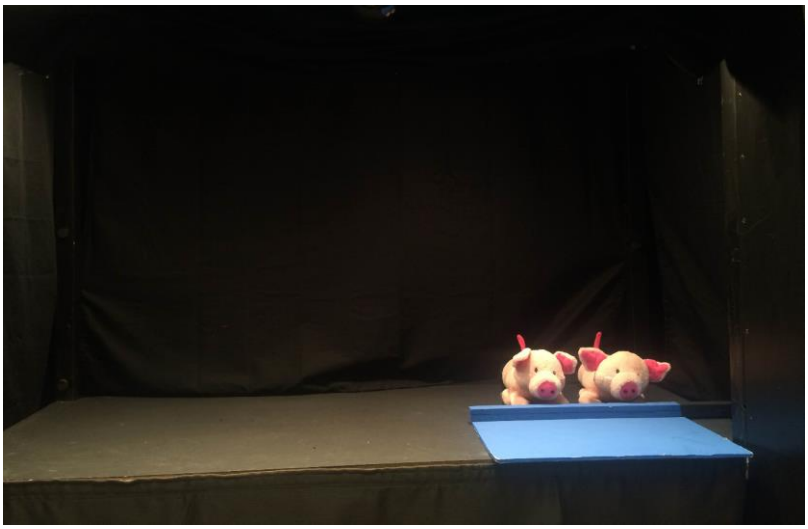
than a property of an individual. Traditional explanations of the process of object individuation in both psychology (Xu & Carey, 1996) and philosophy (Hirsh, 1982) invoke the construct 'kinds' or 'sortal kinds' (Xu, 2005) to explain how people keep track of different objects over time by using conceptual distinctions that apply to *groups* of individuals. While our interpretation of the current results follows this framework, an alternative possibility worth considering is that the current results reflect only infants' attributions of stable sociomoral behaviors to specific individuals without any commitment to kind-based representations. Future research should help clarify this issue by testing whether different looking individuals who share the same sociomoral disposition are more likely to be represented as one object, just as infants are biased to represent an instance of a red mug and a blue glass as a single individual belonging to the kind-category 'cup' (Xu, et. al., 2004).

Overall, the two studies presented here suggest that between 11 and 14 months of age infants conceptualize sociomoral dispositions as a central component in the identity of intentional agents. From a young age we seem biased to represent "good" and "mean" dispositions as relatively stable over time, qualitatively different each other, and based on the attribution of internal properties.

APPENDICES

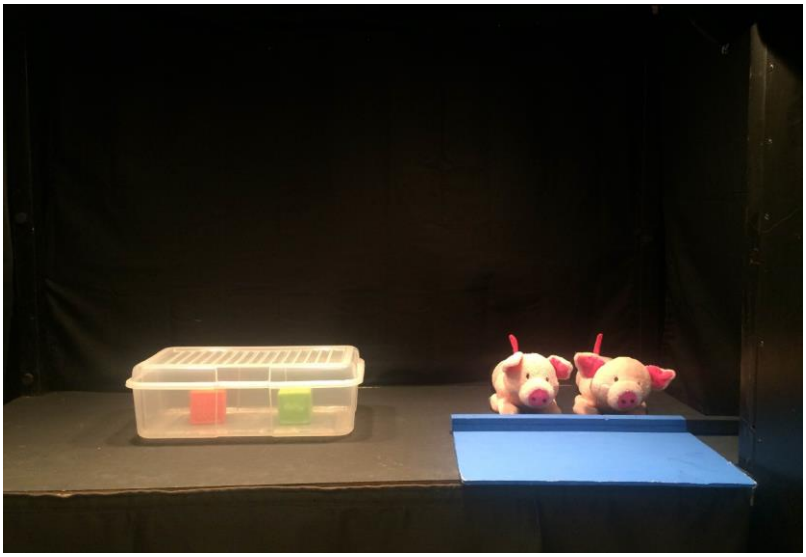
APPENDIX A

PICTURES STIMULI STUDY 1

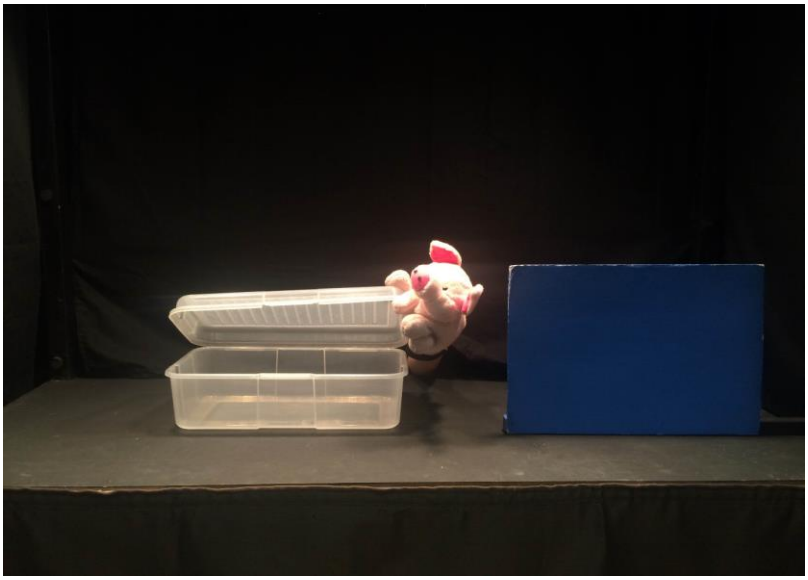


Study 1: Baseline 2 Objects

Study 1: Hindering Action



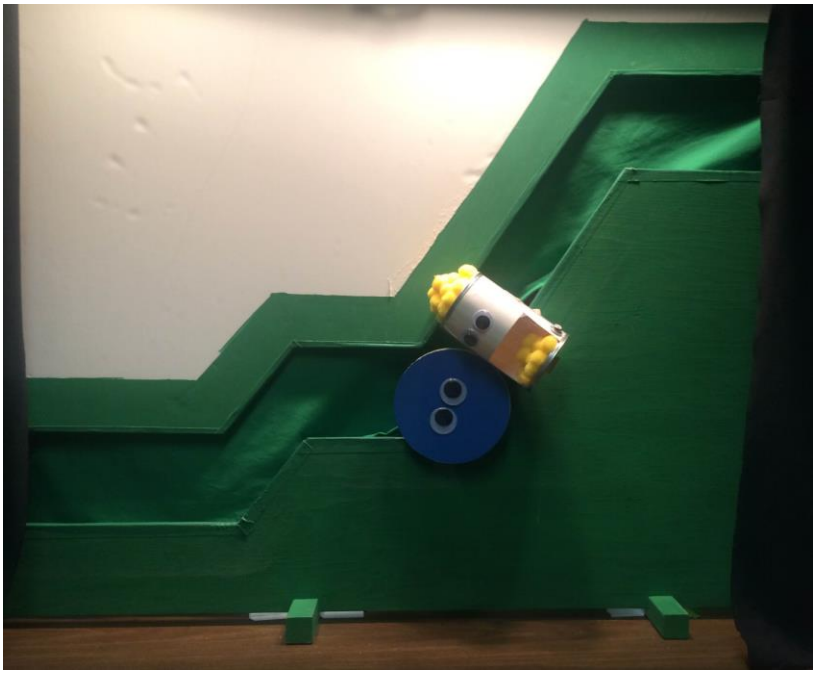
Study 1: Test Trial 2 Objects



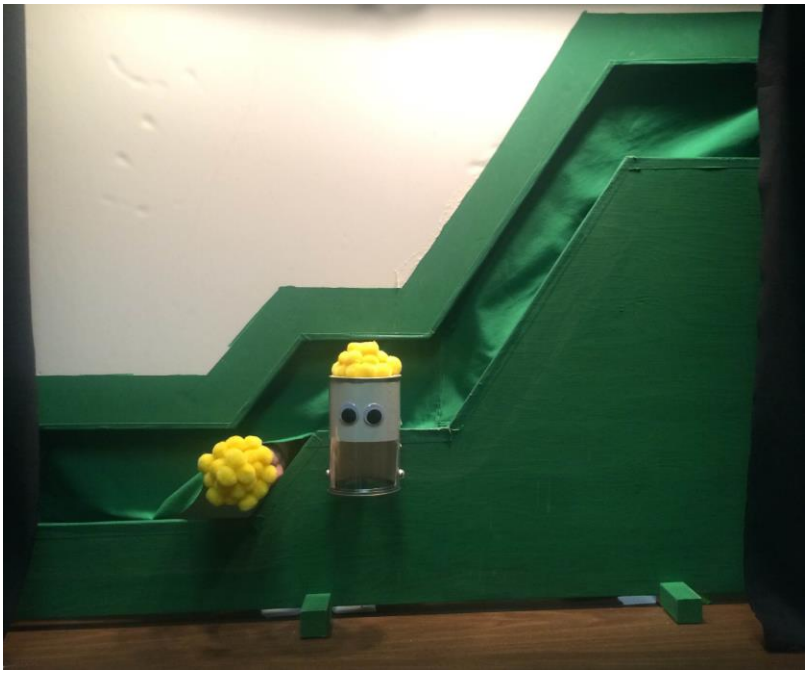
Study 1: Open Condition Experiment 3

APPENDIX B

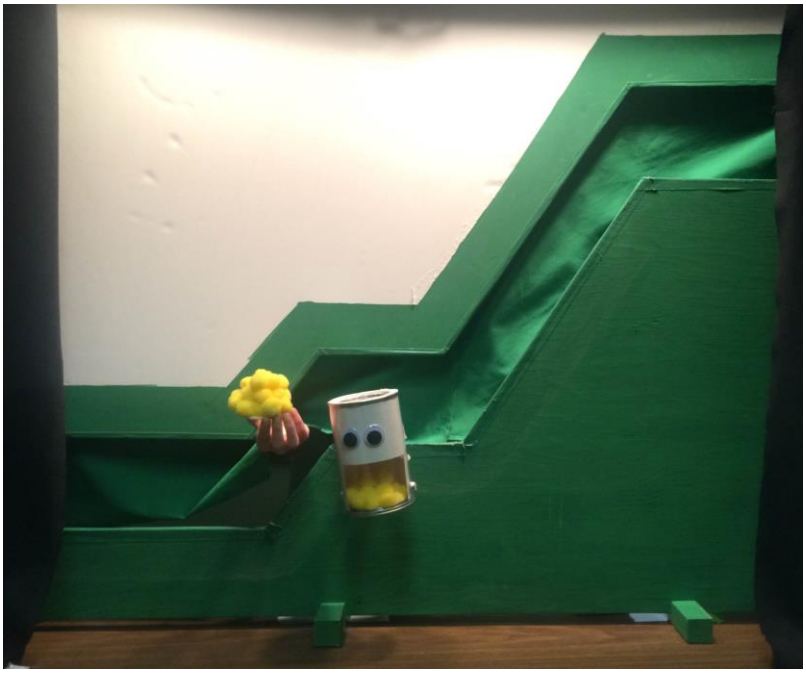
PICTURES STIMULI STUDY 2



Study 2: Hindering Trial



Study 2: Insides Removed Condition



Study 2: Outsides Removed Condition

Study 2: Test Trial

BIBLIOGRAPHY

Aguiar, A., & Baillargeon, R. (1999). 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive Psychology*, 39, 116–157.

Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: the role of causal status in determining feature centrality. *Cognition*, 69, 135-178.

Ahn, W., Gelman, S. A., Amsterlaw, J. A., Hohenstein, J., & Kalish, C. W. (2000). Causal status effect in children's categorization. *Cognition, 76*, B35-B43.

Ahn, W., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., Coley, J. D., & Shafto, P. (2001). Why essences are essential in the psychology of concepts: commentary on Strevens. *Cognition, 82*, 59-69.

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., & Lee, K. (2010). Categorization, categorical perception, and asymmetry in infants' representation of face race. *Infant and Child Development*, 13, 4, 553-564.

Atran, S. (1998). Folk biology and the anthropology of science. Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 31, 547–609.

Atran, S. (1999). Itzaj Maya folkbiological taxonomy: Cognitive universals and cultural particulars. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 233–284). Cambridge, MA: The MIT Press.

Baillargeon, R., Stavans, M., Wu, D., Gertner, Y., Setoh, P., Kittredge, P., & Bernard, A. (2012). Object individuation and physical reasoning in infancy: an integrative account, *Language Learning and Development, 8*, 4-46.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, 64, 3-26.

Baldwin, D. A., Markman, E. M., & Melartin, E. M. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development*, 64, 711–728.

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. (2006). Nature and nurture in own-race face processing. *Psychological Science*, 17, 159–163.

Barrett, H. C., Laurence, S., & Margolis, E. (2008). Artifacts and original intent: A cross cultural perspective on the design stance. *Journal of Cognition and Culture*, 8, 1–22.

Birnbaum, D., Deeb, I., Segall, G., Ben-Eliyahu, A., & Diesendruck, G. (2010). The development of social essentialism: The case of Israeli children's inferences about Jews and Arabs. *Child Development*, 81, 757–777.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, 60, 1-29.

Bloom, P. (1998). Theories of artifact categorization. *Cognition*, 66, 87-93.

Bloom, P., & Markson, L. (1998). Intention and analogy in children's naming of pictorial representations. *Psychological Science*, 9, 200–204

Bonatti, L., Frot, E., Zangl, R., & Mehler, J. (2002). The human first hypothesis: identification of conspecifics and individuation of objects in young infants. *Cognitive Psychology*, 44, 388–426.

Booth, A. E. (2008). The cause of infant categorization? *Cognition*, 106, 984–993.

Booth, A. E., &Waxman, S. R. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, 38, 948–957.

Booth, A.E., Schuler, K, & Zajicek, R. (2010). Specifying the role of function in infant categorization. *Infant Behavior and Development*, 33, 4, 672-684.

Boyer, P., (1993). Pseudo-natural kinds. In P. Boyer, (ed.), *Cognitive Aspects of Religious Symbolism*, Cambridge: Cambridge University Press, pp. 121-141.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, 59, 247–274.

Cacchione, T., Schaub, S., & Rakoczy, H. (2013). Fourteen-month-old infants infer the continuous identity of objects on the basis of non-visible causal properties. *Developmental Psychology*, 49(7), 1325-1329.

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

Carey, S. (2009). *The Origin of Concepts*. USA: Oxford University Press.

Carey, S., & Spelke, E.S. (1994). Domain specific knowledge and conceptual change. In L. Hirschfeld & S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture.* Cambridge: Cambridge University Press, 169-200.

Carey, S., & Xu, F. (2001). Infants knowledge of objects: Beyond object-files and object tracking. *Cognition, 80,* 179-213.

Cimpian, A., & Markman, E. M. (2011). The generic/non-generic distinction influences how children interpret new information about social others. *Child Development*, 82, 471–492.

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Sciences*, 7, 173–179.

Csibra, G. & Gergely, G. (2007). 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, 124, 60-78.

Csibra, G. & Shamsudheen, R. (2015). Nonverbal generics: Human infants interpret objects as symbols of object kinds. *Annual Review Psychology*, 66, 689-710.

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science*, 1, 255–259.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13, 148–153.

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of 'pure reason' in infancy. *Cognition*, 72, 237-267.

Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: On the deceptive determinism of DNA. *Psychological Bulletin*, 137, 800–818.

Dewar, K. & Xu, F. (2007). Do 9-month-old infants expect distinct words to refer to kinds? *Developmental Psychology*, 43, 1227-1238.

Dewar, K., & Xu, F. (2009). Do early nouns refer to kinds or distinct shapes? Evidence from 10-month-old infants. *Psychological Science*, 20(2), 252-257.

Diesendruck, G. & Eldror, E. (2011). What children infer from social categories. *Cognitive Development*, 26, 118-126.

Diesendruck, G. (2001). Essentialism in Brazilian children's extensions of animal names. *Developmental Psychology*, 37, 1, 49-60.

Diesendruck, G., & haLevi, H. (2006). The role of language, appearance, and culture in children's social category-based induction. *Child Development*, 77, 539–553.

Diesendruck, G., & Peretz, S. (2013). Domain differences in the weights of perceptual and conceptual information in children's categorization. *Developmental Psychology*, 49, 2383–2395.

Diesendruck, G., Gelman, S.A. & Lebowitz, K. (1998). Conceptual and linguistic biases in children's word learning. *Developmental Psychology*, 34, 823-839.

Diesendruck, G., Goldfein-Elbaz, R., Rhodes, M., Gelman, S. A., & Neumark, N. (2013). Cross-cultural differences in children's beliefs about the objectivity of social categories. *Child Development,* 84, 1906–1917.

Diesendruck, G. & Lindenbaum, T. (2009). Self-protective optimism: Children's biased beliefs about the stability of traits. *Social Development, 18, 4,* 946-961.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences*, 12, 248–253

Ereshefsky, M. (2010). What is wrong with the new biological essentialism. *Philosophy of Science*, 775, 674-685

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. *Memory and Cognition*, 31, 2, 199-214.

Feigenson, L. & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants' manual search. *Developmental Science*, 6, 568-584.

Futo, J., Teglas, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition, 117*, 1-8.

Gelman, R. (1990). First principles organize attention to relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science*, 14, 79-106.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, 20, 65 – 95.

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

Gelman, S. A. (2004). Psychological essentialism in children. *TRENDS in Cognitive Sciences*, 8:9, 404-409.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology,* 26, 796–804.

Gelman, S. A., & Ebeling, K. S. (1998). Shape and representational status in children's early naming. *Cognition*, 66, B35–B47.

Gelman, S. A., & Koenig, M. A. (2003). Theory-based categorization in early childhood. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion* (pp. 330–359). New York: Oxford University Press.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23, 183–209.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understanding of the non-obvious. *Cognition, 38*, 213–244.

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

Gergely, G., & Jacob, P., (2012). Reasoning about Instrumental and Communicative Agency in Human Infancy. In J. B. Benson (Serial Ed.) & F. Xu & T. Kushnir (Vol. Eds.), *Rational Constructivism in Cognitive Development* (pp. 59–94). Elsevier Inc.: Academic Press

Gergely, G., Nádasdy, Z., Csibra, G., & Biıró, S. (1995).Taking the intentional stance at 12 months of age. *Cognition*, 56, 165–193.

German, T. P., & Johnson, S. C. (2002). Function and the origins of the design stance. *Journal of Cognition and Development*, 3, 279–300.

Gil-White, F. J. (2001). Are ethnic groups biological "species" to the human brain? Essentialism in our cognition of some social categories. *Current Anthropology*, 42, 515–554.

Gottfried, G. M., & Gelman S. A. (2005). Developing domain-specific causal-explanatory frameworks: The role of insides and immanence. *Cognitive Development*, 20, 137–158.

Graham, S. A., & Diesendruck, G. (2010). Fifteen-month-old infants attend to shape over other perceptual properties in an induction task. *Cognitive Development*, 25, 111-123.

Graham, S. A., & Kilbreath, C. S. (2007). It's a Sign of the Kind: Gestures and words guide infants' inductive inferences. *Developmental Psychology*, 43(5), 1111-1123.

Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). 13-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development*, 75, 409–427.

Graham, S., Keates, J., Vukatana E., & Khu, M. (2013). Distinct labels attenuate 15-month-olds' attention to shape in an inductive inference task. *Frontiers in Psychology*, 3, 1-8.

Gutheil, G., & Rosengren, K. (1996). A rose by any other name: Preschoolers' understanding of individual identity across name and appearance changes. *British Journal of Developmental Psychology*, 14, 477–498.

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

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450, 557–560.

Hamlin, J.K. (2013a). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, 22, 3, 186 – 193

Hamlin, J.K. (2013b). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, 128, 3, 451 – 474.

Hamlin, J.K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: Infants prefer those who harm dissimilar others. *Psychological Science*, 24, 589-594.

Hampton, J. A. (2001). The role of similarity in natural categorization. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.), *Similarity and categorization* (pp. 13-28). Oxford: Oxford University Press.

Haslam, N. and Ernst, D. (2002) Essentialist beliefs about mental disorders. *Journal of Social and Clinical Psychology*, 21, 628–644

Haslam, N. O. (1998). Natural kinds, human kinds, and essentialism. *Social Research*, 65, 291-314.

Haslam, N. O., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, 39, 113-127.

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*, 30, 1661 – 1673

Hatano, G., & Inagaki, K. (1994). Young children's naïve theory of biology. *Cognition*, 50, 171–188.

Haukioja, J. (2014). On deriving essentialism from the theory of reference. *Philosophical Studies*, 172, 8, 2141-2151.

Hayes BK, & Rehder B. (2012). The development of causal categorization. *Cognitive Science*, 36, 6, 1102–28.

Heyman, G. D., & Gelman, S. A. (2000a). Beliefs about the origins of human psychological traits. *Developmental Psychology*, 36, 665-678.

Heyman, G. D., &Gelman, S. A. (2000b). Preschool children's use of novel predicates to make inductive inferences about people. *Cognitive Development*, 15, 263-280.

Heyman,G. D.,&Gelman,S. A. (1998).Young children use motive information to make trait inferences. *Developmental Psychology*, 34, 310–321.

Hirsch, E. (1982). *The concept of identity*. New York: Oxford University Press.

Hirschfeld, L. A. (1996). *Race in the making: Cognition, culture, and the child's construction of human kinds.* Cambridge, MA: MIT Press.

Hirschfeld, L. A., & Gelman, S. A. (1997). What young children think about the relation between language variation and social difference. *Cognitive Development*, 12, 213–238.

Howard, L.H., Henderson, A.M., Carrazza, C. & Woodward, A. (2015). Infants' and young children's imitation of linguistic in-group and out-group informants. *Child Development*, 86, 1, 259-275.

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

Jylkka, J. W. (2008). Theories of natural kind term reference and empirical psychology. *Philosophical Studies*, 139, 153–169.

Jylkka, J., Railo, H., & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology*, 22, 37–60.

Kalish, C.W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition*, 23, 335-353.

Kalish, C.W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition*, 30, 340-352.

Keil, F. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Keil, F. Greif M. & Kerner, R. S. (2007). A world apart: How concepts of the constructed world are different in representation and in development. In E. Margolis & S. Laurence (Eds.), *Creations of the Mind: Theories of artifacts and their representation* (pp. 212-230). USA: Oxford University Press.

Keil, F.C. (1991). The Emergence of Theoretical Beliefs as Constraints on Concepts. In S. Carey and R. Gelman (Eds.), *The Epigenesis of Mind: Essays on Biology and Cognition* (pp. 237-256). Earlbaum

Keil, F.C. (1995). The Growth of Causal Understandings of Natural Kinds: Modes of Construal and the Emergence of Biological Thought. In. A. Premack and D. Sperber (Eds.), *Causal Cognition*. Oxford: Oxford University Press.

Kingo, O. S., & Krojgaard, P. (2011). Object manipulation facilitates kind-based object individuation of shape-similar objects. *Cognitive Development*, 26, 2, 103-111.

Kinzler, K. D., & Dautel, J. B. (2012). Children's essentialist reasoning about language and race. *Developmental Science*, 15, 131–138.

Kinzler, K.D., & Spelke, E. (2011). Do infants show social preferences for people differing in race? *Cognition*, 119, 1–9.

Kinzler, K.D., Dupoux, E., & Spelke, E.S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 12577–12580.

Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Journal of Cognitive Neuroscience*, 14, 5, 402–408.

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

Luo, Y & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science, 16, 8*, 601-608.

Malt, B. C. (1994). Water is not H2O. *Cognitive Psychology*, 27, 41-70.

Malt, B. C., & Sloman, S. A. (2007). Category essence or essentially pragmatic? Creator's intention in naming and what's really what. *Cognition*, 105, 615–648

Malt, B.C., & Johnson, E.C. (1992). Do artifact concepts have cores? *Journal of Memory and Language*, 31, 195–217.

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development*, 1, 1, 3–36.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, 8, 291–318.

Mandler, J. M., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition*, 59, 307–335.

Mandler, J.M. (2003). Conceptual categorization. In D.H. Rakison & L.M. Oakes (Eds.), *Early Category and Concept Development* (pp. 103-130). USA: Oxford University Press.

Mandler, J.M. (2004a). *The foundations of Mind: Origins of conceptual thought*. USA: Oxford University Press.

Mandler, J.M. (2004b). Thought before language. *Trends in Cognitive Science*, 8, 508–513.

Mandler, J.M., &McDonough, L. (1998). Studies in inductive inference in infancy. *Cognitive Psychology*, 37, 60–96.

Markson, L., Diesendruck, G., & Bloom, P. (2008). The shape of thought. *Developmental Science*, 11, 204-208.

Mascaro O, & Csibra G. (2012). Representation of stable social dominance relations by human infants. *Proceedings of the National Academy of Sciences*. USA 109 (18): 6862–67.

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition*, 78, 1–26.

McDonough, L., & Mandler, J.M. (1998). Inductive generalization in 9- and 11-month-olds. *Developmental Science*, 1, 227–232

Medin, D. L. & Ortony, A. (1989). Psychological essentialism. In *S.* Vosniadou & A. Ortony (Eds), *Similarity and Analogical Reasoning.* New York: Cambridge University Press.

Meunier, B., & Cordier, F. (2009). The role of feature type and causal status in 4-5-years-old children's biological categorizations. *Cognitive Development*, 24, 34-48.

Newman, G. E., Herrmann, P., Wynn, K., & Keil, F. C. (2008). Biases towards internal features infants' reasoning about objects. *Cognition*, 107, 420-432.

Pauen, S. (2002a). Evidence for knowledge-based category discrimination in infancy. *Child Development*, 73(4), 1016–1033.

Pauen, S. (2002b). The global-to-basic level shift in infants' categorical thinking: First evidence from a longitudinal study. *International Journal of Behavioral Development*, 26(6), 492–499.

Plunkett, K., Hu, J.F., & Cohen, L.B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106, 2, 665-681.

Pomiechowska, B., Tatone, D., & Csibra, G. (2016). Infants individuate agents involved in dyadic social relations through the principle of relational consistency. Poster presented at the Conference on Cognitive Development, Budapest, Hungary.

Powell, L.J., & Spelke, E.S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences,* 110, 3965-3972.

Prasada, S., Hennefield, L., & Otap, D. (2012). Conceptual and linguistic representations of kinds and classes. *Cognitive Science,* 36, 1224–1250.

Premack, D. (1990). The Infants Theory of Self-Propelled Objects. *Cognition, 36(1),* 1-16.

Prentice, D.A., & Miller, D.T. (2006). Essentializing differences between women and men. *Psychological Science*, 17, 129– 135.

Prentice, D.A., & Miller, D.T. (2007). Psychological essential- ism of human categories. *Current Directions in Psychological Science*, 16, 202–206.

Putnam, H. (1975). *Mind, Language, and Reality*. New York: Cambridge University Press.

Rochat, P., Striano, T., & Morgan, R. (2004). Who is doing what to whom? Young infants' developing sense of social causality in animated displays. *Perception, 33,* 3, 355-369.

Quine, W. V. (1969). Natural kinds. In W. V. Quine, *Ontological relativity and other essays* (pp. 114-138). New York: Columbia University Press.

Quinn, P. C. (2002). Category representation in infants. *Current Directions in Psychological Science, 11,* 66-70.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perceptio*n, 22, 463–475.

Quinn, P., Eimas, P., & Tarr, M. (2001). Perceptual categorization of cat and dog silhouettes by 3-to 4-month infants. *Journal of Experimental Child Psychology*, 79, 1, 78-94.

Quinn, P., Yahr, J., Kuhn, A., Slater, A., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, 31, 1109–1121

Quinn, P.C. & Eimas, P.D. (1996). Perceptual organization and categorization in young infants. In C. Rovee-Collier & L.P. Lipsitt (Eds.). *Advances in Infancy Research* (Vol. 11, pp. 1-36). Norwood, NJ: Ablex.

Rakison, D.H. & Cicchino, J.B. (2009). Development of inductive inference in infancy. In S.P. Johnson (Ed.), *Neoconstructivism. The new science of cognitive development* (pp. 233-251). USA: Oxford University Press.

Rakison, D.H. (2003). Part, motion, and the development of the animate-inanimate distinction in infancy. In D.H. Rakison & L.M. Oakes (Eds.), *Early Category and Concept Development* (pp. 159-192). USA: Oxford University Press.

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, 27, 709-748.

Rehder, B., & Kim, S. W. (2009). Classification as diagnostic reasoning. *Memory & Cognition*, 37, 715–729.

Rhodes, M., & Brickman, D. (2011). The influence of competition on children's social categories. *Journal of Cognition and Development*, 12, 194–221.

Rhodes, M., & Gelman, S. A. (2009a). A developmental examination of the conceptual structure of animal, artifact, and human social categories across two cultural contexts. *Cognitive Psychology*, 59, 244–274.

Rhodes, M., & Gelman, S.A. (2009b). Five-year-olds beliefs about the discreteness of category boundaries for animals and artifacts. *Psychonomic Bulletin & Review, 16, 5*, 920-924.

Rhodes, M., Leslie, S., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, 109, 13526–13531.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). Cambridge: Cambridge University Press.

Rips, L. J. (2011). *Lines of Thought: Central concepts in cognitive psychology*. US: Oxford University Press.

Rostad, K., Yott, J., & Poulin-Dubois, D. (2012). Development of categorization in infancy: Advancing forward to the animate/inanimate level. *Infant Behavior and Development*, 35, 3, 584-595.

Scholl, B. J., & Leslie, A. M. (1999). Explaining the infant's object concept: Beyond the perception/cognition dichotomy. In E. Lepore & Z. Pylyshyn (Eds.), *What is cognitive science?* (pp. 26-73). Oxford: Blackwell.

Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences*, USA, 110 (40), 15937-15942.

Sloman, S. (2005). *Causal Models: How we think about the world and its alternatives*. Oxford University Press.

Sloman, S. A., & Malt, B. C. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. In H. E. Moss & J. A. Hampton (Eds.), *Conceptual representation* (pp. 563-582). Hove, U.K.:

Spelke, E. S., Kestenbaum, R., Simons, D. J., & Wein, D. (1995). Spatiotemporal continuity, smoothness of motion, and object identity in infancy. *British Journal of Developmental Psychology*, 13, 113-142.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition*, 74, 149-175.

Surian, L., & Caldi, S. (2010). Infant's individuation of agents and inert objects. *Developmental Science*, 13, 1, 143-150.

Taborda-Osorio, H. & Cheries, E. (2015). Infants' agent individuation: it's what's on the inside that counts. *Cognition*, under review.

Taylor, M. G. (1996). The development of children's beliefs about social and biological aspects of gender categories. *Child Development*, 67, 1555–1571.

Taylor, M.G., Rhodes, M., & Gelman, S.A. (2009). Boys will be boys, cows will be cows: Children's essentialist reasoning about gender categories and animal species. *Child Development*, 80, 461-481.

Thomsen, L., Frankenhuis, W., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: preverbal infants mentally represent social dominance. *Science, 331*, 477-480.

Tremoulet, P., Leslie, A., & Hall, D. G. (2000). Infant individuation and identification of objects. *Cognitive Development*, 15, 499–522.

Van de Walle, G. A., Carey, S., & Prevor, M. (2000). Bases of object individuation in infancy: Evidence from manual search. *Journal of Cognition and Development*, 1, 3, 249-280.

Ware, E. A., & Booth, A. E. (2010). Form follows function: The role of artifact function in the development of the shape bias. *Cognitive Development*, 25, 124–137.

Waxman, S.R. and Grace, A.D. (2012). Developing gender- and race-based categories in infants: Evidence from 7- and 11-month-olds. In G. Hayes & M. Bryant (Eds.), *Psychology of Culture*. In *Psychology of Emotions, Motivations and Actions: Focus on Civilizations and Cultures* Series. Hauppauge, NY: Nova Science Publishers.

Welder, A. N., & Graham, S. A. (2001). The influence of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, 72, 1653–1673.

Welder, A. N., & Graham, S. A. (2006). Infants' categorization of novel objects with more or less obvious features. *Cognitive Psychology*, 52, 57–91.

Wilson, R. A., Barker, M. J., & Brigandt, I. (2007). When Traditional Essentialism Fails: Biological Natural Kinds. *Philosophical Topics*, 35, 1-2, 189-215.

Woodward, A. L., & Hoyne, K. L. (1999). Infants' learning about words and sounds in relation to objects. *Child Development*, 70, 65–77.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358, 749-750.

Wynn, K. (2008). Some innate foundations of social and moral cognition. In P. Carruthers, S. Laurence & S. Stich (Eds.), *The Innate Mind: Foundations and the Future*. Oxford: Oxford University Press

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85, 223–250.

Xu, F. (2005). Categories, kinds, and object individuation in infancy. In L. Gershkoff-Stowe and D. Rakison (eds.), *Building object categories in developmental time* (pp. 63-89). Papers from the 32nd Carnegie Symposium on Cognition. New Jersey: Lawrence Erlbaum

Xu, F. (2007). Sortal concepts, object individuation, and language. *Trends in Cognitive Sciences*, *11*, 400-406.

Xu, F., & Baker, A. (2005). Object individuation in 10-month- old infants using a simplified manual search method. *Journal of Cognition and Development*, 6, 3, 307–323.

Xu, F., & Carey, S. (1996). Infant's metaphysics: The case of numerical identity. *Cognitive Psychology,* 30, 111-153.

Xu, F., Carey, S., & Quint, N. (2004). The emergence of kind-based object individuation in infancy. *Cognitive Psychology*, 49, 155–190.

Xu, F., Carey, S., & Welch, J. (1999). Infants' ability to use object kind information for object individuation. *Cognition*, 70, 137-166.

Yuill, N. (1992). Children's conception of traits. *Human Development*, 35, 265-279.

Yuill, N., & Pearson, A. (1998). The development of bases for trait attribution: Children's understanding of traits as causal mechanisms based on desire. *Developmental Psychology*, 34, 574-586.