Précis for "The inner life of goals: costs, rewards, and commonsense psychology."

Julian Jara-Ettinger

1. Introduction

Theories of decision-making have been at the heart of psychology since the field's inception, but only recently has the field turned to the study of how humans think humans make decisions. When we watch someone make a choice, we explain it in terms of their goals, preferences, personalities, and moral beliefs. This capacity – our commonsense psychology – is the cognitive foundation of human society. It lets us share what we have and know, with those from whom we expect the same in return, and it guides how we evaluate those who deviate from our expectations.

The representations and inferential power underlying commonsense psychology trace back to early childhood – before children begin kindergarten, and often even in infancy. Nonetheless, major theoretical questions remain unresolved. What computations underlie our commonsense psychology, and to what extent are they specific to the social domain? Are there a small number of general principles by which humans reason about and evaluate other agents, or do we instead learn a large number of special case rules and heuristics? To what extent is there continuity between the computations supporting commonsense psychology in infancy and later ages? Is children's social-cognitive development a progressive refinement of a computational system in place from birth, or are there fundamentally new computational principles coming into play?

This thesis presents a hypothesis that offers answers to each of these questions, and provides a unifying framework in which to understand the diverse social-cognitive capacities we see even in young children. I propose that human beings, from early infancy, interpret others' intentional actions through the lens of a naïve utility calculus: we assume that others act to maximize utilities -the rewards they expect to obtain relative to the costs they expect to incur. I argue that this principle is at the heart of social cognition, possibly from early infancy. It can be made precise computationally and tested quantitatively. Embedded in a Bayesian framework for reasoning under uncertainty, and supplemented with other knowledge children have about the physical and psychological world (e.g. knowledge about objects, forces, action, perception, goals, desires, and beliefs), the naïve utility calculus supports a surprisingly wide range of core social-cognitive inferences. It persists stably in some form through adulthood, guiding the development of social reasoning even as children's thinking about others undergoes significant growth.

Figure 1 illustrates some of the basic social intuitions that the naïve utility calculus aims to explain. These examples illustrate the principles that drive our intuitions in a wide range of situations in which intentional agents of any sort (child, adult, animated ball) interact with each other and move toward, reach for, or manipulate objects. I focus on behaviors such as those shown in Figure 1 where even young children can immediately grasp the costs and rewards involved. The naïve utility calculus likely applies to more abstract situations as well, but its application may be complex in ways I do not consider here (e.g., cases where cultural norms are in play). Although I focus on intentional behavior (as opposed to habits, reflexes, accidents, etc.)

some of the most revealing choices are decisions not to act; the naïve utility calculus aims to account for these as well.

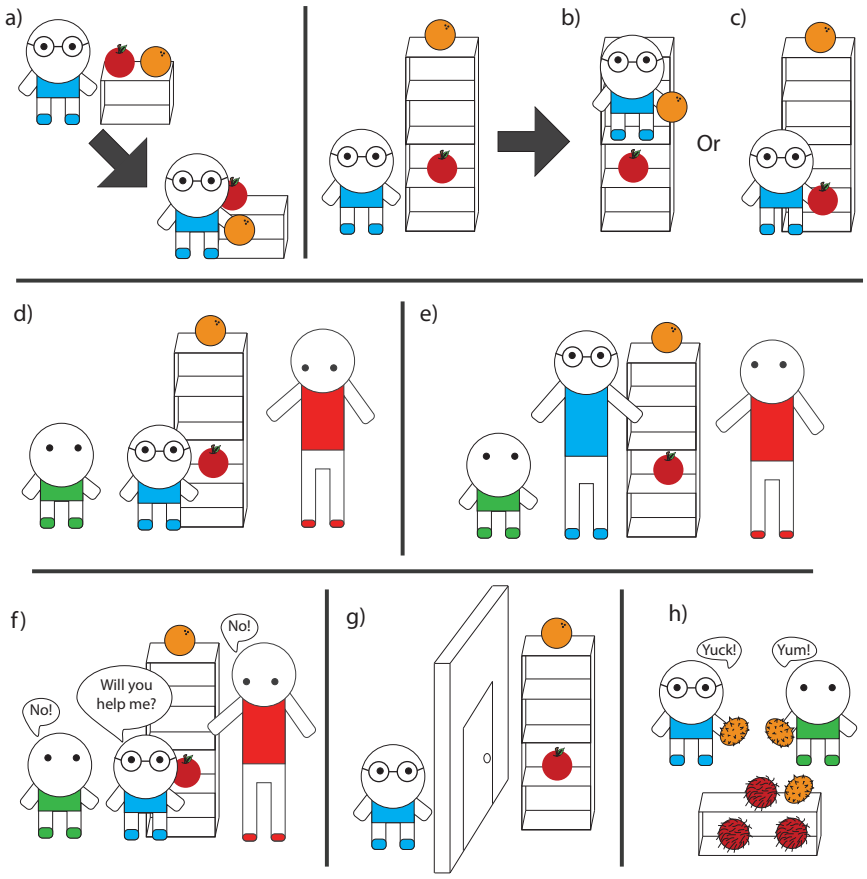

Figure 1. The logic of costs and rewards underlying our commonsense psychology. (a) If the blue agent (the protagonist) chooses the orange over the apple, her immediate goal is clearly the orange, but how confident are you that she prefers oranges in general to apples? (b) If the orange were high on the top shelf and the agent climbs up to get it, would you become more confident she prefers oranges in general? (c) What if she had chosen the apple instead? Does this indicate any strong preference for apples? (d) If the blue agent wants the orange from the top shelf, whom should she ask for help? (e) If she is the tallest person in the room, is it still appropriate for her to ask for help? (f) If both the red and green agent refuse to help, are they equally mean or is the red one meaner? (g) If the blue agent cannot see the shelf and says she is going to get the orange, are you confident she won't change her mind? (h) If both agents choose kiwanos over rambutans, but one says "Yum!" and the other says "Yuck!" after tasting it, who's more likely to have never tasted the fruits before?

Critically, the naïve utility calculus is not a scientific account of how people act; it is a scientific account of people's intuitive theory of how people act. The naïve utility calculus does not require that agents actually compute and maximize fine-grained expected utilities in order to be a useful guide in many everyday social situations.

This thesis has three goals. The first goal is to evaluate how the naïve utility calculus relates to other theories of goal-directed action understanding, and to evaluate its explanatory power

with respect to existing empirical data. The second goal is to test key predictions of the naïve utility calculus in children. If we fundamentally reason about others through the assumption of utility maximization, then this reasoning should be already at work in early childhood. The third goal is to develop a formal computational account of the naïve utility calculus which I compare against quantitative judgments from adult participants. This allows me to ensure the theory's precision and to test its fine-grained predictions.

2. The naïve utility calculus

Chapter 2 presents the naïve utility calculus, sketches out its main qualitative predictions (summarized in Figure 2), and discusses how the theory explains basic social intuitions that are already at work in infancy (Jara-Ettinger, Gweon, Schulz, & Tenenbaum; 2016).

The naïve utility calculus naturally produces the expectation that agents should behave efficiently (Gergely & Csibra, 2003; Scott & Baillargeon, 2013). If agents maximize utilities, they must also necessarily minimize costs. As such, the naïve utility calculus is consistent with the "teleological stance" (Gergely & Csibra, 2003). However, the naïve utility goes beyond the teleological stance, as it also explains how agents choose which goals to pursue. As such, the naïve utility calculus can explain a goal-directed action understanding. The naïve utility calculus also explains a wide range of studies about how toddlers and infants infer preferences based on spatial and statistical information (Kushnir, Xu, & Wellman, 2010; Gweon, Tenenbaum, & Schulz, 2010; this idea is tested quantitatively in Chapter 7), how they make social evaluations (Jara-Ettinger, Tenenbaum, & Schulz, 2015; tested directly in Chapter 5), and their intuitions about pedagogical events (e.g., Bridgers, Jara-Ettinger, & Gweon, 2016).

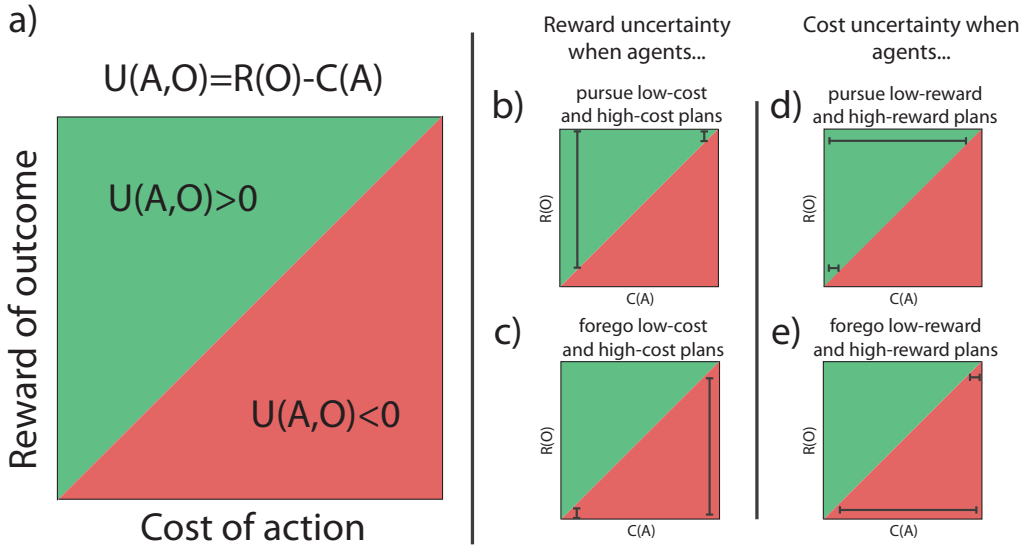

Figure 2. Utility maximization in scenarios where costs depend on actions (*C(A)*) and rewards depend on outcomes (*R(O)*). (a) Qualitative plot of how the decision to act reflects the total utility. The utility is positive (*U(A,O)>0*) when the rewards outweigh the costs (*R(O)>C(A)*) and negative otherwise. (b) If an agent pursues a low-cost plan, a wide range of rewards could have produced a positive utility, thus low cost actions do not reveal much about the agent's reward. However, if the agent pursues a high-cost plan, then the reward must also be high. (c) The structure of these inferences flips when the agent refuses to pursue a plan. If agent refuses to pursue a low cost plan, the reward must also be low.

By contrast, a wide range of rewards is consistent with refusing to pursue a costly plan, such that if the agent refuses to pursue a high cost plan, her refusal is not very informative about her rewards. (d) The implications are parallel when we infer costs given reward knowledge. Low rewards only motivate action when the costs are low, while high rewards motivate action under a wide range of costs; thus the pursuit of a goal when rewards are low is more informative about an agent's costs (that they are probably low). (e) If an agent foregoes a low reward we may be uncertain about the costs of acting, but if she foregoes a high reward we can be more certain that the costs are high.

Beyond its explanatory power, the naïve utility calculus makes new untested predictions. If we assume that agents act to maximize utilities, the naïve utility calculus predicts that, for observers, some actions are more informative than others (see also Lucas et al, 2014). If we watch an agent incur a low cost to obtain a reward, we cannot be sure about the magnitude of this reward, because both low and high rewards can produce positive utilities. By contrast, if we watch an agent incur a high cost to obtain a reward, we can be sure that the reward was high (otherwise the utility would have been negative; Fig 2b). These inferences flip in the case of inaction. If an agent chooses not to pursue a low-cost plan, we can be sure that the reward must have been very low. However, if an agent chooses not to purse a high-cost plan, this does not imply that reward was low (Fig 2c). The nature of these inferences is parallel when we know the agents' rewards and are trying to infer the costs (Fig 2d and 2e).

3. Breaking into the infer life of goals

Chapter 3 presents four experiments that directly test if children have a naïve utility calculus (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015). In Experiment 1 we show that when five-year-olds learn an agent's costs and choices, they infer a reward function through the assumption that agents maximize utilities. Children watched a puppet choose crackers over cookies when both items were equidistant, but cookies over crackers when the cookies were closer. If children equate choice with preference, they should think the puppet likes crackers and cookies equally; instead, our results show that children integrate the puppet's choices with cost information and recognize that the puppet prefers crackers (i.e., the item chosen when the costs were matched). In Experiment 2 we show that when children observe agents' choices whose rewards are known, they infer a cost function that guarantees that the agents maximized their utilities. We showed children one puppet who liked crackers more than cookies and another puppet who liked them both equally. We then put the cookies on a low box and the crackers on a high box and both puppets chose the cookies. When asked which puppet couldn't climb, children chose the puppet with the strong preference even though neither puppet even attempted to climb. Further experiments also showed that children can design interventions that will best allow them to infer other agents' competence (Experiments 3 and 4).

4. Action under uncertainty

For the naïve utility calculus to work, it is critical to understand that agents maximize the utilities they *expect* to obtain (see Figures 1g and 1h). As such, agents who are ignorant or wrong about the costs and rewards of are more likely to make poor choices and to change their minds in light of new experiences (and conversely, agents who make poor choices and change their minds are more likely to be ignorant or wrong about their costs and rewards). Chapter 4 presents four experiments showing that children share these intuitions (Jara-Ettinger, Lydics=, Tenenbaum, &

Schulz, 2015; Jara-Ettinger, Floyd, Tenenbaum, & Schulz, *under review*). In Experiment 1 we introduced four- and five-year-olds to two puppets, both of whom reached for and chose a rambutan over an kiwano. The two puppets took a bit from their fruit and one of them said "yuck" (or, in Experiment 2, she changed her mind). Children were asked which puppet had never seen these fruits before. Children successfully identified the naïve agent. Conversely, we found that if children learn which agent is knowledgeable and which agent is naïve, they can infer which puppet is more likely to dislike their choice, and to revise it (Experiments 3 and 4).

5. Social Reasoning

If the assumption of utility maximization is fundamental to social cognition, it should also underlie how we reason about social goals. The naïve utility calculus has especially interesting implications in social evaluations. When an agent refuses to help, it is crucial to determine if they refused because of a lack of motivation (which is morally reprehensible) or because of a lack of competence (which is not). The naïve utility calculus predicts that we should judge more harshly someone who refuses to incur a low cost to help compared to someone who refuses to incur a high cost to help (because the former reveals a lack of motivation but the latter does not; Figures 1f and 2c). In Chapter 5 we show that toddlers share this intuition (Jara-Ettinger, Tenenbaum, & Schulz, 2015). We showed two-year-old children two puppets making a toy play music; one puppet was able to make the toy play music on the first try (low cost) while the other took several attempts (high cost). At baseline, toddlers preferred to play with the more competent agent and judged him to be nicer. When both puppets refused to help a parent activate the toy, toddlers continued to prefer the more competent agent but now judged that the less competent agent was nicer. Consistent with the naïve utility calculus, these results suggest that two-year-olds can infer an agent's motivation to help (her subjective rewards) given information about her costs and, like adults, are more likely to exonerate agents for whom helping is costly than those who are simply unmotivated to be helpful.

6. Formal implementation

In Chapter 6 I present a formal computational account of the naïve utility calculus. This model aims to explain social reasoning at a computational level of analysis, without a commitment to the specific algorithms (Marr, 1982). I formalize the naïve utility calculus as a generative model that probabilistically generates utility-maximizing goals and actions, and I rely on Bayesian inference to invert the model and recover the costs and reward given observable actions. More formally,

$$p(C, R|A) \propto \sum_{g \in G} p(A|g, C, R)p(g|C, R)p(C, R)$$

where C and R are the agents' unobservable costs and rewards, A is the set of observed actions, and G is the set of possible goals that the agent may be pursuing.

This general approach has proved successful in related models of action understanding (Baker, Saxe, & Tenenbaum, 2009; Baker, Jara-Ettinger, Saxe, & Tenenbaum, *under revision*). The naïve utility calculus model predicts with very high accuracy how adults infer other people's

underlying competence and motivation by watching their behavior. Moreover, lesioned models that are conceptually similar, but less powerful, fail to explain participant judgments with the same precision.

7. Unifying early social cognition

In chapter 7 I explore more formally the relation between the naïve utility calculus and children and infants' sensitivity to the sampling process (Jara-Ettinger, Sun, Schulz, & Tenenbaum, *under review*). A growing set of studies suggests that rare choices reveal stronger preferences (e.g., Gweon, Tenenbaum, & Schulz, 2010; Kushnir, Xu, & Wellman, 2010; Wellman, Kushnir, Xu, & Brink, 2016). For instance, if we watched an agent take a red ball (a common choice) from the box in Figure 3a, we would not necessarily conclude that she prefers them over the blue balls. By contrast, if the agent took a blue ball (a rare choice), we would be more likely to infer that the agent prefers them over the red balls. These intuitions can be derived from the naïve utility calculus. If both types of balls are equally rewarding, then an agent who maximizes utilities should simply take whichever ball she comes in contact with first. The less common a ball is, the more likely that the agent will have to incur additional costs in terms of distance, time, and attention to retrieve it. This can be seen if we imagine transforming the sampling scenario (Figure 3a) into a spatial layout (Figure 3b). From an observer's standpoint, if an agent retrieves a rare object, the assumption that they maximize utilities implies that the reward for this rare object must be higher than the reward for the more common object. By contrast, if the agent retrieves a common object, her choice is consistent with having no preference at all (if the rare objects are extremely costly to get, then getting a common type of object is even consistent with the agent preferring the rare object kind).

Through an experiment with adults, I show how the naïve utility calculus model is naturally sensitive to both spatial and statistical features of the environment, and that it fits participant judgments better than models of sensitivity to the sampling process that had been proposed in the past (e.g. Gweon, Tenenbaum, & Schulz, 2010).

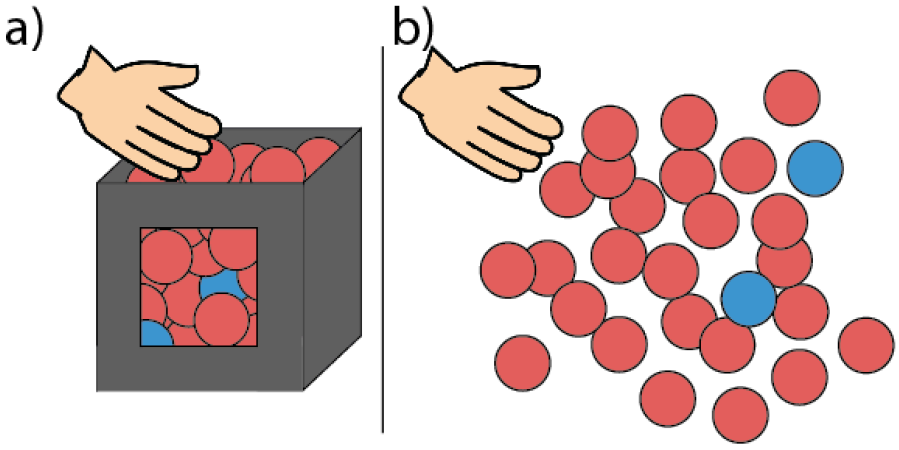

Figure 3. (a) an agent can retrieve a red ball or a blue ball from a box. Intuitively, choosing a blue ball reveals a strong preference whereas choosing a red ball does not. (b) equivalent scenario unfolded spatially. Rare objects are less likely to be the most convenient. As such, an object's rarity correlates with its cost.

8. Conclusions

Altogether, the explanatory power of the naïve utility calculus (Chapters 1 and 2), the developmental evidence in support of it (Chapter 3-5), and the high similarity between our computational model and adult participant judgments (Chapter 7 and 8), provide strong evidence that we fundamentally reason about others through the assumption that others maximize their subjective utilities. Through this assumption, it is possible to work backwards from someone's actions to infer their competence, their preferences, and even their moral culpability.

I have argued that starting early in development, humans interpret other agents' behavior through a naïve utility calculus. The connection between the naïve utility calculus as an account of intuitive decision theory and formal theories of decision- making based on expected utility maximization developed in economics may appear coincidental or simply convenient, but the relation may run deep. As Fritz Heider argued (1958), scientific theories, especially in their early stages, may be grounded on commonsense; what better way to formulate initial hypotheses if not by what we intuitively believe to be true? Heider quotes the physicist Robert Oppenheimer: "...all sciences arise as refinement, corrections, and adaptations of common sense."

Suppose that scientific theories of human decision making, starting with classical utility theory and moving through their descendants in behavioral economics, really began grounded on the common sense theory I discussed here. This view has several implications. First, the reason that our models of common-sense psychology in children look like classical utility theory might be because early economists were, with a different purpose in mind, doing exactly what we do here: formalizing common-sense psychology. Second, our common-sense psychology is, at its core, right. Despite the memorable cases where we fail to understand each other, we get others right more often than not. Even if it fails to account for human decision-making in less ecologically relevant domains (e.g., economic choices in the modern marketplace), the naïve utility calculus, as the first models in utility theory, captures key features of human intentional action in the most basic everyday situations even the youngest children appreciate. And as Heider observed, even when commonsense psychology is wrong with respect with how we make choices, it's still right in an important sense. Our most important everyday choices involve others, and our ability to reason about their own choices influences what we do. This intuitive decision-theory is therefore, by definition, a cornerstone of any scientific theory of human decision making.

Finally, the ways in which people's decision making fails to conform with basic assumptions of classical utility theory, which are often counterintuitive and surprising, are surprising precisely because they go against our common-sense. As such, these surprises may point to features of the naïve theory that we have not yet elucidated. To cite just one salient example, we may overinterpret others' failures to help in a low-cost situation as a sign that they don't value helping us. But maybe our naïve theories do not sufficiently take into account agents' non-optimal planning; they wanted to help but they didn't plan well. Or perhaps our naïve theories oversimplify by assuming we know all the relevant costs (or rewards) even when we don't, or assuming that others' costs are like ours even when they're not; both of these assumptions could lead us to mistake a failure to help as a low-cost refusal even when it isn't. Understanding how our commonsense psychology is oversimplified in these ways could advance not only our understanding of core social cognition as scientists, but also, ultimately, help us better understand each other as human beings.

References

Baker, Jara-Ettinger, Saxe, & Tenenbaum (*under revision*). Bayesian theory of mind: Mental state attribution as inverse planning under uncertainty.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. Children consider others' expected costs and rewards when deciding what to teach.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287-292.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066-9071.

Heider, F. The psychology of interpersonal relations. Psychology Press, 1958.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not So Innocent Toddlers' Inferences About Costs and Culpability. *Psychological science*, *26*(5), 633-640.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14-23.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589-604.

Jara-Ettinger, J., Sun, F., Schulz, L. E., & Tenenbaum, J. B., (*under review*). The principle of efficiency unifies spatial and statistical routes to preference.

Jara-Ettinger, J., Lydic, E., Tenenbaum, J., & Schulz, L. E. (2015). Beliefs about desires: Children's understanding of how knowledge and preference influence choice. Proceedings of the 37th Annual Conference of the Cognitive Science Society.

Jara-Ettinger J., Floyd S., Tenenbaum, J. B., & Schulz L.E. (*under review*). Children believe that agents maximize expected utilities.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134-1140.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

Marr, D. (1982). Vision. ”W. H. Freeman and Company, San Francisco, CA.

Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychological science*, 0956797612457395.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*.